Doing the Right Things: Leaders Wanted … Apply Within Sounding the Call to Arms

Greg Hutto & Jim Simpson
Ops Analysts Test Wings
Air Armament Center
Eglin AFB, Florida
Gregory.hutto@eglin.af.mil

NASA Statistical Engineering Symposium 5 May 2011

Structure

- If this DOE stuff is so good … why do I struggle?
- Outline of a story to convince our leaders
- Equipping leaders with the right questions to ask
- Summary & Questions

If all this DOE Stuff is so good … why do I struggle?

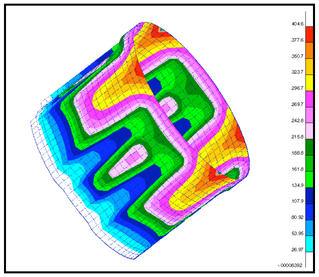
Deming and the VP – May be Apocryphal, but True …

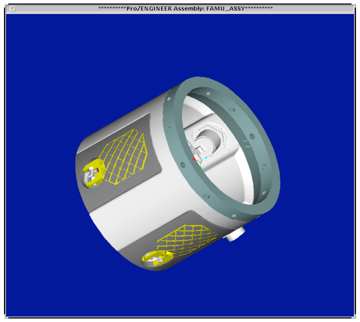
"Learning is not compulsory . . . neither is survival.“

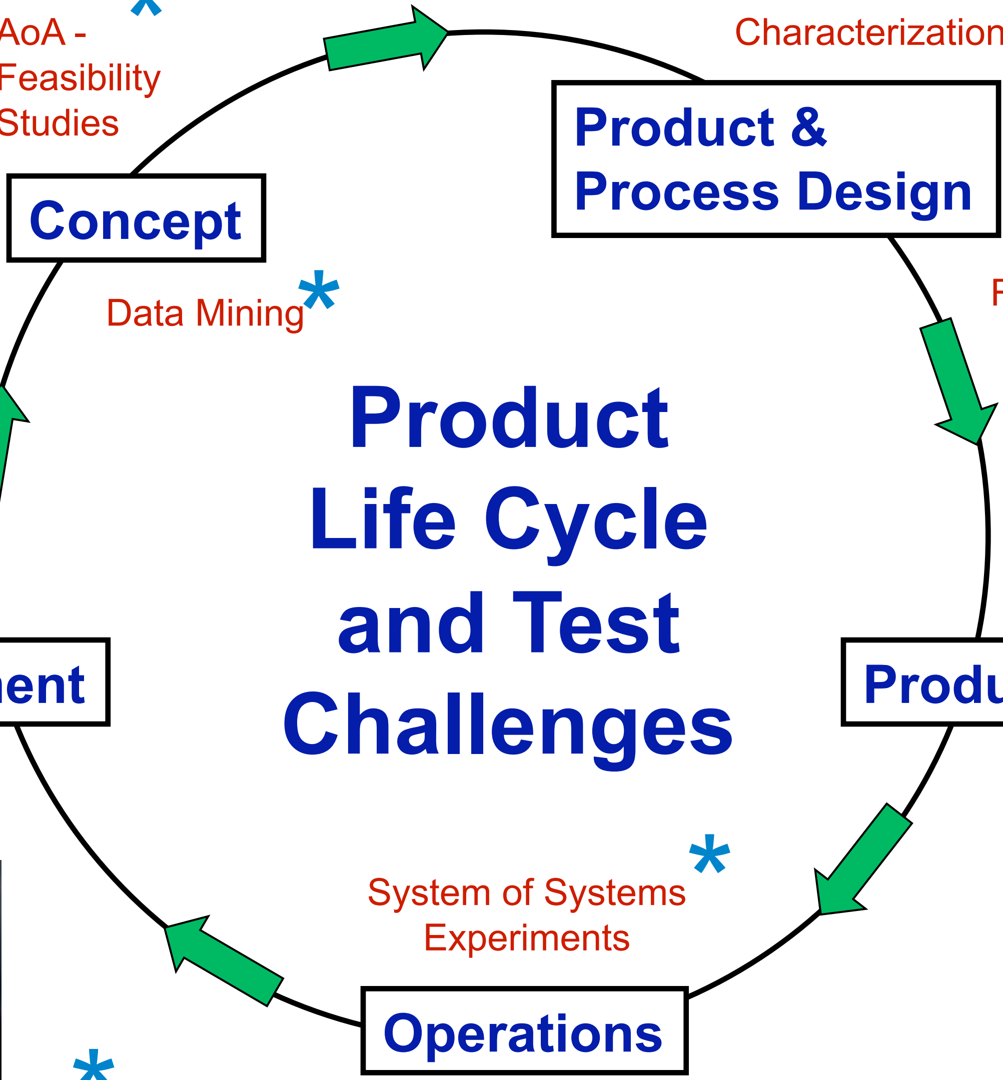
"It is not enough to do your best; you must know *what* to do, and then do your best."

-- W. Edwards Deming

October 14, 1900 – December 20, 1993







Product Life Cycle and Test Challenges

Design Variable Screening *

AoA - Feasibility Studies *

Product Characterization *

Requirements - Quality Function Deployment (QFD)

Data Mining *

Robust Product Design * *

Performance Mapping

Product Optimization *

Measurement Systems Analysis *

System of Systems Experiments *

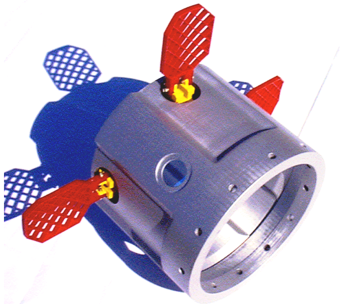
Reliability and Sustainment Testing *

* Reliability

Maintainability

Availability

Serviceability



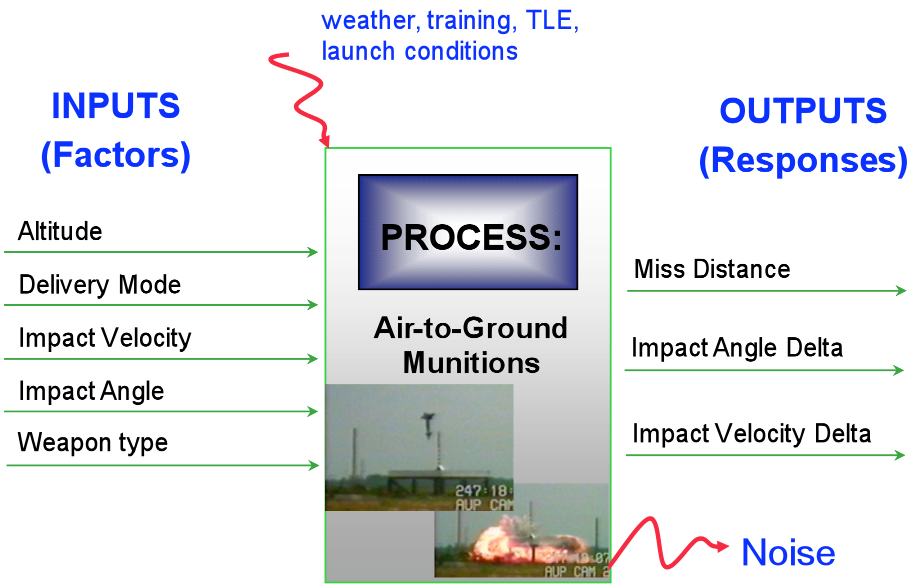
* Examples just this conference

Systems Engineering Employ Many Simulations of Reality

		Simulation of Reality					
Acq Phase		M&S		Hardware		System/Flight Test	
Reqt'ts Dev		Warfare					
	AoA		Physics				
Concepts				HWIL/SIL	Captive	Subsystem	Prototype
	Risk Reduction						
EMD							Prod Rep
	Prod & Mfr						
Sustain							Production

- At each stage of development, we conduct experiments
  - Ultimately – how will this device function in service (combat)?
  - Simulations of combat differ in fidelity and cost
  - Differing goals (screen, optimize, characterize, reduce variance, robust design, trouble-shoot)
  - Same problems – distinguish truth from fiction: What matters? What doesn’t?

What are Statistically Designed Experiments?



weather, training, TLE, launch conditions

INPUTS (Factors)

Altitude

Delivery Mode

Impact Velocity

Impact Angle

Weapon type

PROCESS:

Air-to-Ground Munitions

OUTPUTS (Responses)

Miss Distance

Impact Angle Delta

Impact Velocity Delta

Noise

- Purposeful, systematic changes in the inputs in order to observe corresponding changes in the outputs
- Results in a mathematical model that predicts system responses for specified factor settings

$$\text{Responses} = f\left(\text{Factors}\right) + \varepsilon$$

Case DT/OT: B-1 Radar TLE Accuracy Characterization (2001)

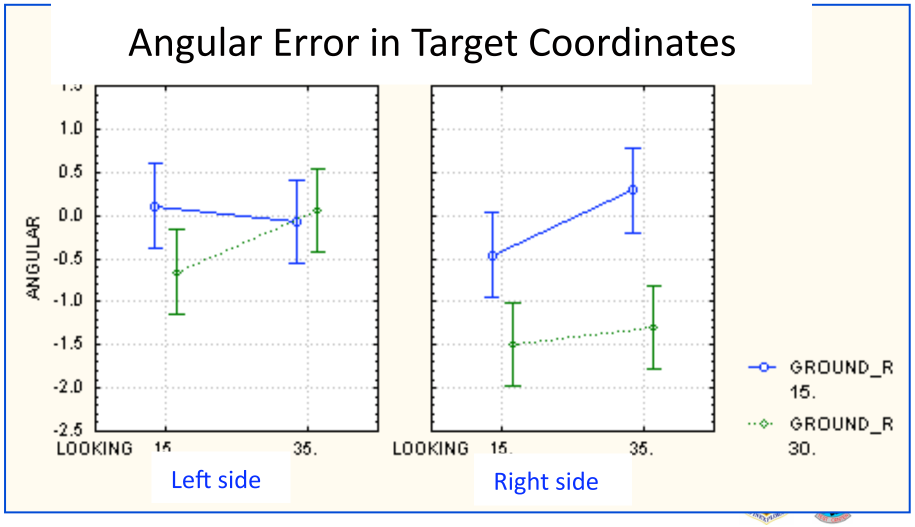
Problem:

- Is B-1B APQ-164 monopulse SAR mode for targeting accurate enough for JDAM?
- Are tail numbers similar? Target types?
- Bottom line: self-target JDAM?
- 7 sorties flown with mixed results
  -100's of measurements "as available"

DOE Approach:

- Variables include
  - Side of A/C, angle off nose
  - Range, type of target
  - Two tail numbers
- Responses include TLE, mil error
- Compare to specified radar accuracy
- Single 2-ship sortie

Results: Similar accuracy across volume, tail



Case: DT HWIL GWEF Large Aircraft IR Hit Point Prediction

IR Missile C-5 Damage

Test Objective:

- IR man-portable SAMs pose threat to large aircraft in current AOR
- Dept Homeland Security desired Hit point prediction for a range of threats needed to assess vulnerabilities
- Solution was HWIL study at GWEF (ongoing)

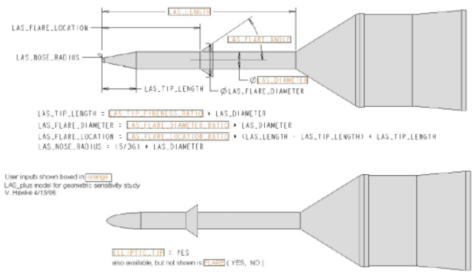
DOE Approach:

- Aspect – 0-180 degees, 7each
- Elevation – Lo,Mid,Hi, 3 each
- Profiles – Takeoff, Landing, 2 each
- Altitudes – 800, 1200, 2 each
- Including threat – 588 cases
- With usual reps nearly 10,000 runs
- DOE controls replication to min needed

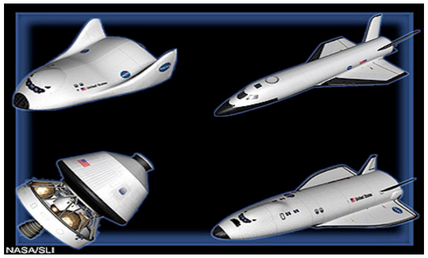
Results:

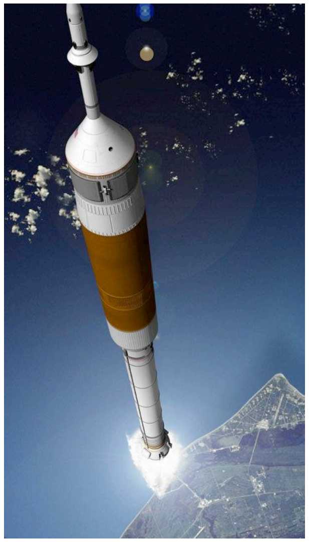
- Revealed unexpected hit point behavior
- Process highly interactive (rare 4-way)
- Process quite nonlinear w/ 3rd order curves
- Reduced runs required 80% over past
- Possible reduction of another order of magnitude to 500-800 runs

Case 11: CFD for NASA CEV



LAS_LENGTH
LAS_FLARE_LOCATION
LAS_FLARE_ANGLE
LAS_NOSE_RADIUS
ØLAS_DIAMETER
LAS_TIP_LENGTH
ØLAS_FLARE_DIAMETER
LAS_TIP_LENGTH = LAS_TIP_FINENESS_RATIO * LAS_DIAMETER
LAS_FLARE_DIAMETER = LAS_FLARE_DIAMETER_RATIO * LAS_DIAMETER
LAS_FLARE_LOCATION = LAS_FLARE_LOCATION_RATIO * (LAS_LENGTH - LAS_TIP_LENGTH) + LAS_TIP_LENGTH
LAS_NOSE_RADIUS = (5/36) * LAS_DIAMETER
User inputs shown boxed in orange
LAS_plus model for geometric sensitivity study
V. Hawke 4/13/06
ELLIPTIC_TIP : YES
also available, but not shown is FLARE (YES, NO)





Test Objective:

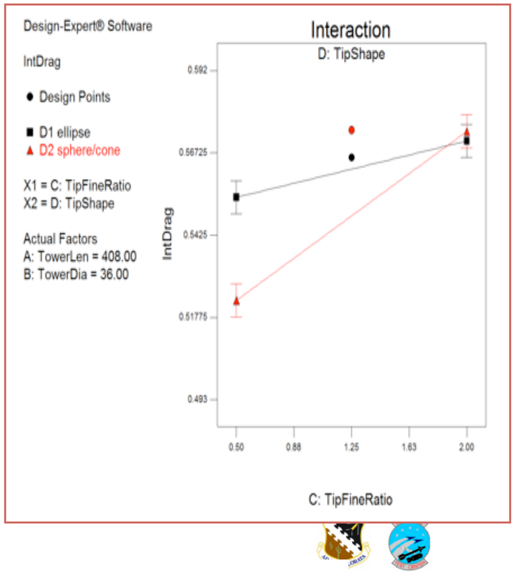
- Select geometries to minimize total drag in ascent to orbit for NASA's new Crew Exploration Vehicle (CEV)
- Experts identified 7 geometric factors to explore including nose shape
- Down-selected parameters further refined in following wind tunnel experiments

DOE Approach:

- Two designs – with 5 and 7 factors to vary
- Covered elliptic and conic nose to understand factor contributions
- Both designs were first order polynomials with ability to detect nonlinearities
- Designs also included additional confirmation points to confirm the empirical math model in the test envelope

Results:

- Original CFD study envisioned 1556 runs
- DOE optimized parameters in 84 runs – 95%!
- ID'd key interaction driving drag



Source: **A Parametric Geometry CFD Study Utilizing DOE** *Ray D. Rhew, Peter A. Parker, NASA Langley Research Center,* **AIAA 2007 1616**

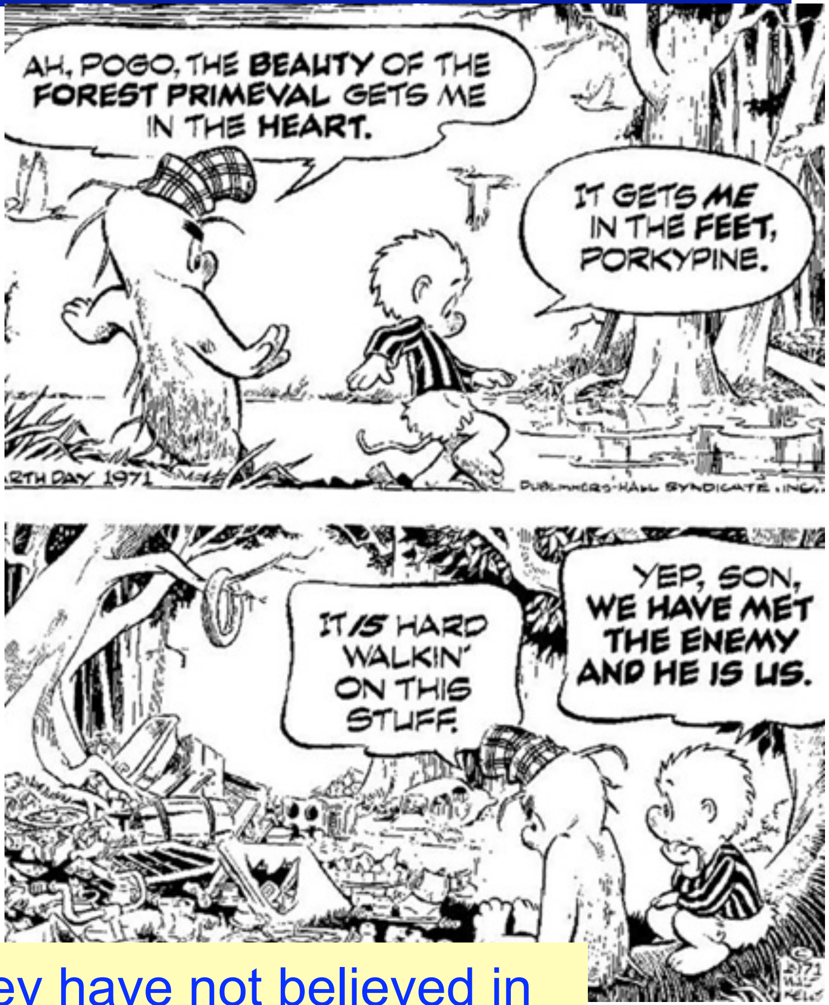
So … *why* aren’t *all* experiments well-designed?

- Summary of three projects:
  - 1 mission when 7 couldn’t answer the question
  - Cut runs from 5000 replicates to 500
  - CFD Trials reduced from 1920 to 84
- Many such outstanding success stories
- We know how to teach & mentor practitioners
- Experts can be hired and groomed
- We have plenty of good software tools, texts

"We have met the enemy and he is … Us!
-- Pogo circa 1971

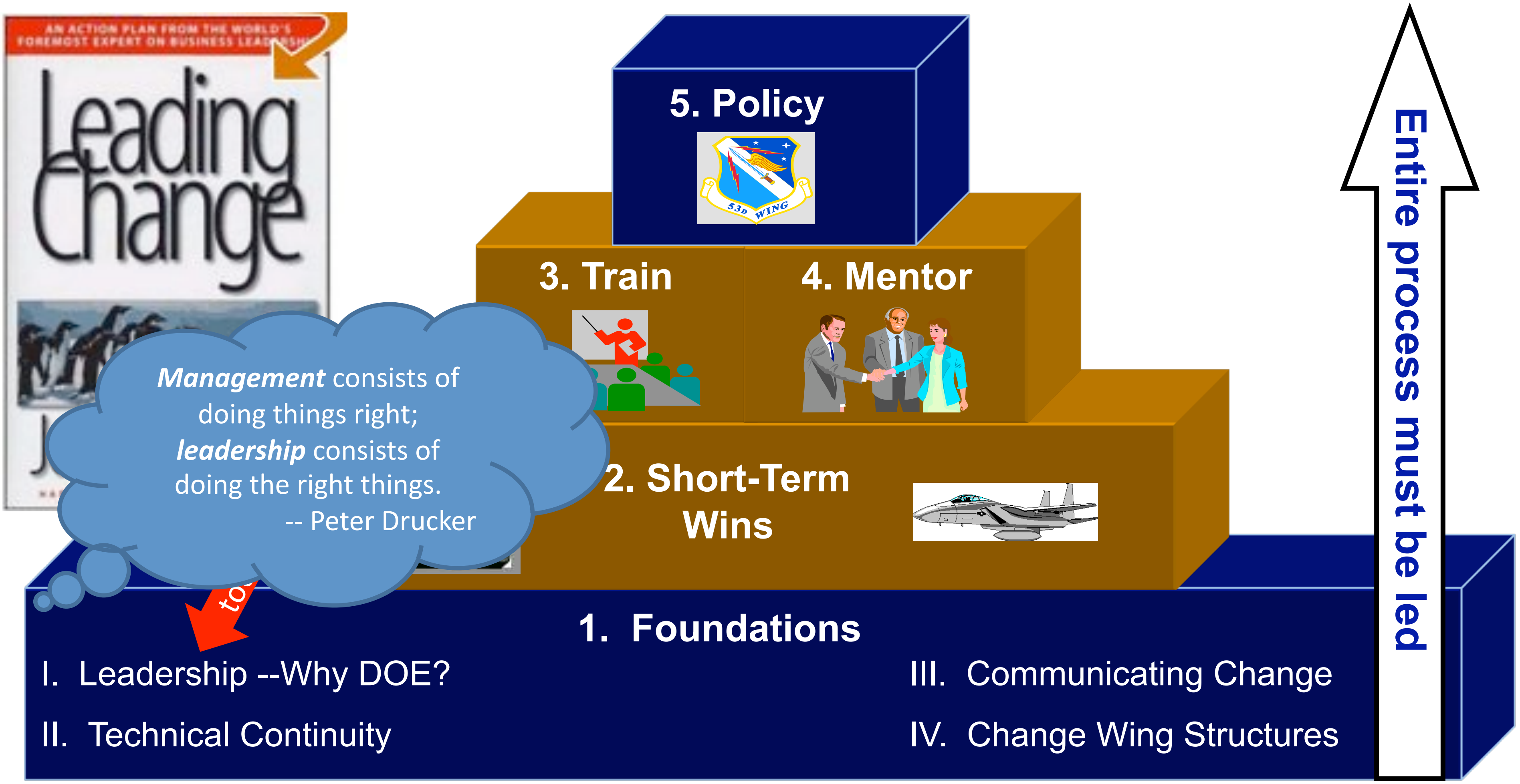
- It is us…
- A Job Story circa 1990-2000
- "Leadership From Below"

  -- Col T.S. Hutto 1933-1998


AH, POGO, THE BEAUTY OF THE FOREST PRIMEVAL GETS ME IN THE HEART.
IT GETS ME IN THE FEET, PORKYPINE.
EARTH DAY 1971
PUBLISHERS-HALL SYNDICATE, INC.
IT IS HARD WALKIN' ON THIS STUFF.
YEP, SON, WE HAVE MET THE ENEMY AND HE IS US.

"But how can people call on him if they have not believed in him? How can they believe in him if they have not heard his message? How can they hear if no one tells the Good News? "
-- Paul (0063, Romans 10.14)

Five Steps to Implementation


AN ACTION PLAN FROM THE WORLD'S FOREMOST EXPERT ON BUSINESS LEADERSH
Leading Change
to

Management consists of doing things right;
leadership consists of doing the right things.
-- Peter Drucker

5. Policy

3. Train

4. Mentor

2. Short-Term Wins

1. Foundations

I. Leadership --Why DOE?

II. Technical Continuity

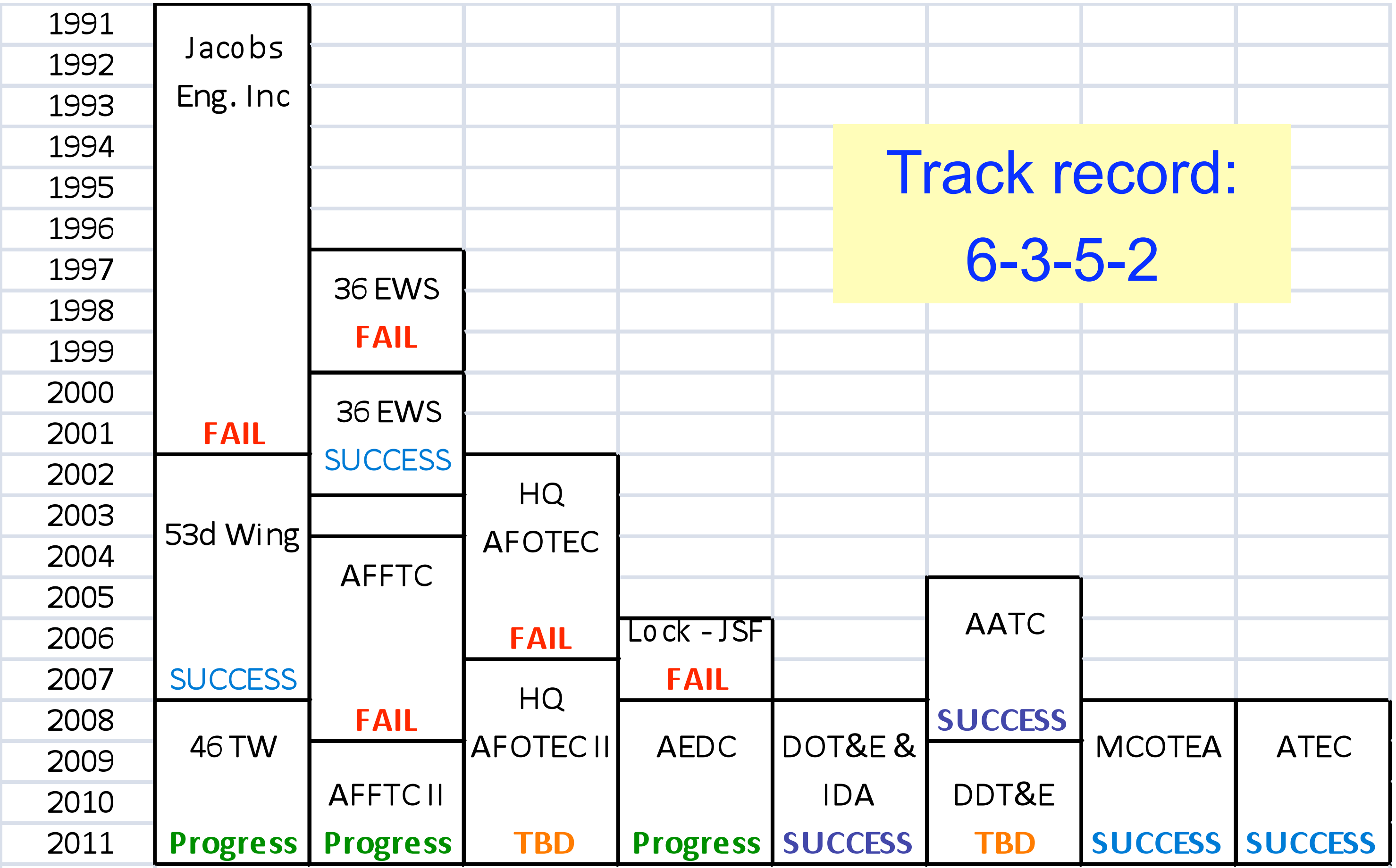
III. Communicating Change

IV. Change Wing Structures

Entire process must be led

"Because **management** deals mostly with the **status quo** and **leadership** deals mostly with **change**, in the next century we are going to have to try to become much more skilled at creating leaders." -- Dr. John Kotter

Telling the "Why?" Story … It is not easy or guaranteed of success



Track record:
6-3-5-2

FAIL = Pockets of success but exec not organize/train/equip/measure to sustain

PROGRESS = Efforts to organize/train/equip/hire and accountability by senior exec

TBD = Encouraging engagements with staff, executives

SUCCESS = Exec establishes accountability, resources, hires, policy. Majority DOE

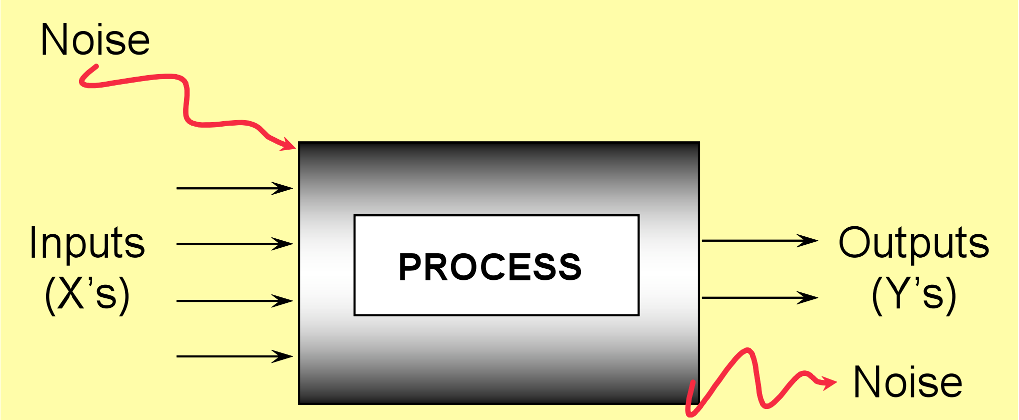
Why DOE? One Slide…
DOE Gives Scientific Answers to Four Fundamental Test Challenges

Four Challenges faced by any test

1. *How many? Depth of Test* – effect of test size on uncertainty
2. *Which Points? Breadth of Testing* – spanning the vast employment battlespace
3. *How Execute? Order of Testing* – insurance against "unknown-unknowns"
4. *What Conclusions? Test Analysis* – drawing objective, scientific conclusions while controlling noise

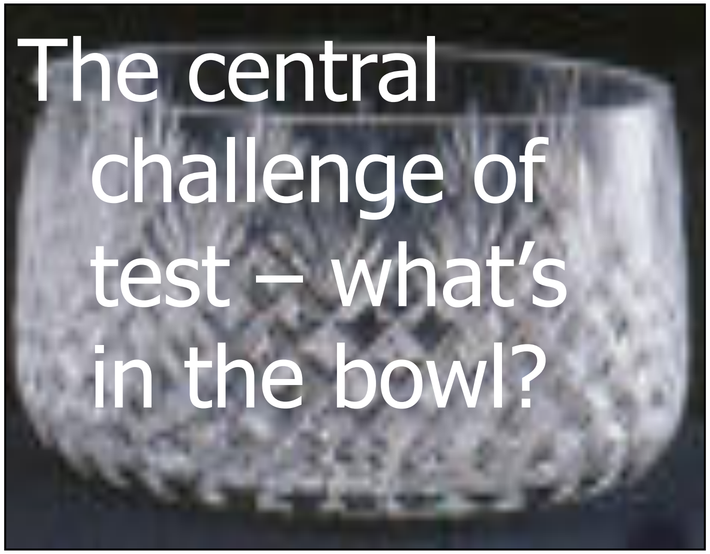
DOE effectively addresses all these challenges!

In our short time today, address primarily #1 and #2.



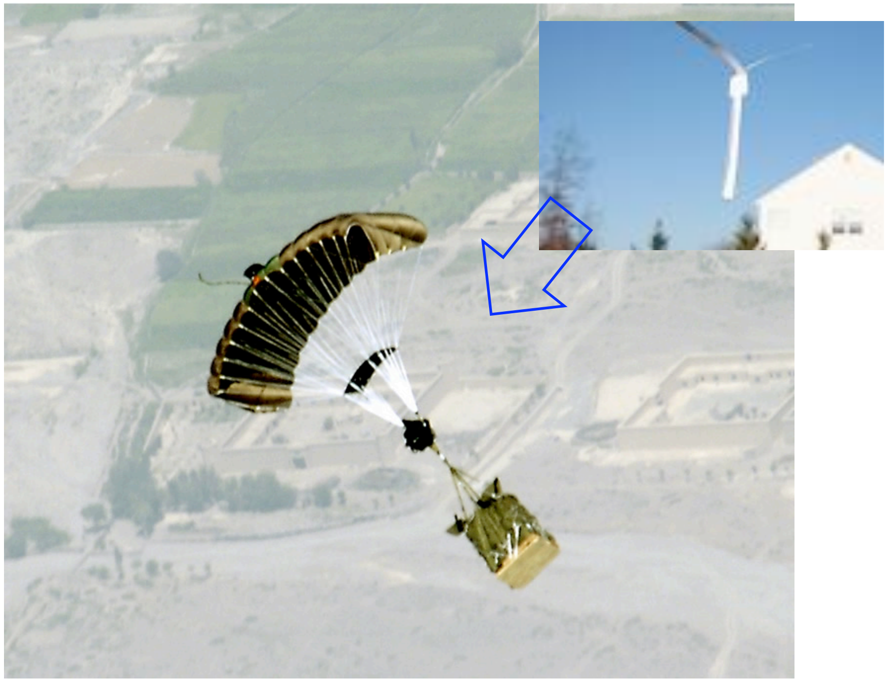
Question #1 … How Many?

- In all our testing – we reach into the bowl (reality) and draw a sample of JPADS performance
- Consider an "80% JPADS"
  - Suppose a required 80% P(Arrival)
  - Is the Concept version acceptable?
- We don't know in advance which bowl God hands us …
  - The one where the system *works* or,
  - The one where the system *doesn't*



The central challenge of test – what's in the bowl?

Example:
Precision Air Drop System



The dilemma for airdropping supplies has always been a stark one. High-altitude airdrops often go badly astray and become useless or even counter-productive. Low-level paradrops face significant dangers from enemy fire, and reduce delivery range. Can this dilemma be broken?

A new advanced concept technology demonstration shows promise, and is being pursued by U.S. Joint Forces Command (USJFCOM), the U.S. Army Soldier Systems Center at Natick, the U.S. Air Force Air Mobility Command (USAF AMC), the U.S. Army Project Manager Force Sustainment and Support, and industry. The idea? Use the same GPS-guidance that enables precision strikes from JDAM bombs, coupled with software that acts as a flight control system for parachutes. JPADS (the Joint Precision Air-Drop System) has been combat-tested successfully in Iraq and Afghanistan, and appears to be moving beyond the test stage in the USA… and elsewhere.

Capability:

Assured SOF re-supply of material

Requirements:

Probability of Arrival
Unit Cost $XXXX
Damage to payload
Payload
Accuracy
Time on target
Reliability …

- Just when you think of a good class example – they are already building it!
- 46 TS – 46 TW Testing JPADS

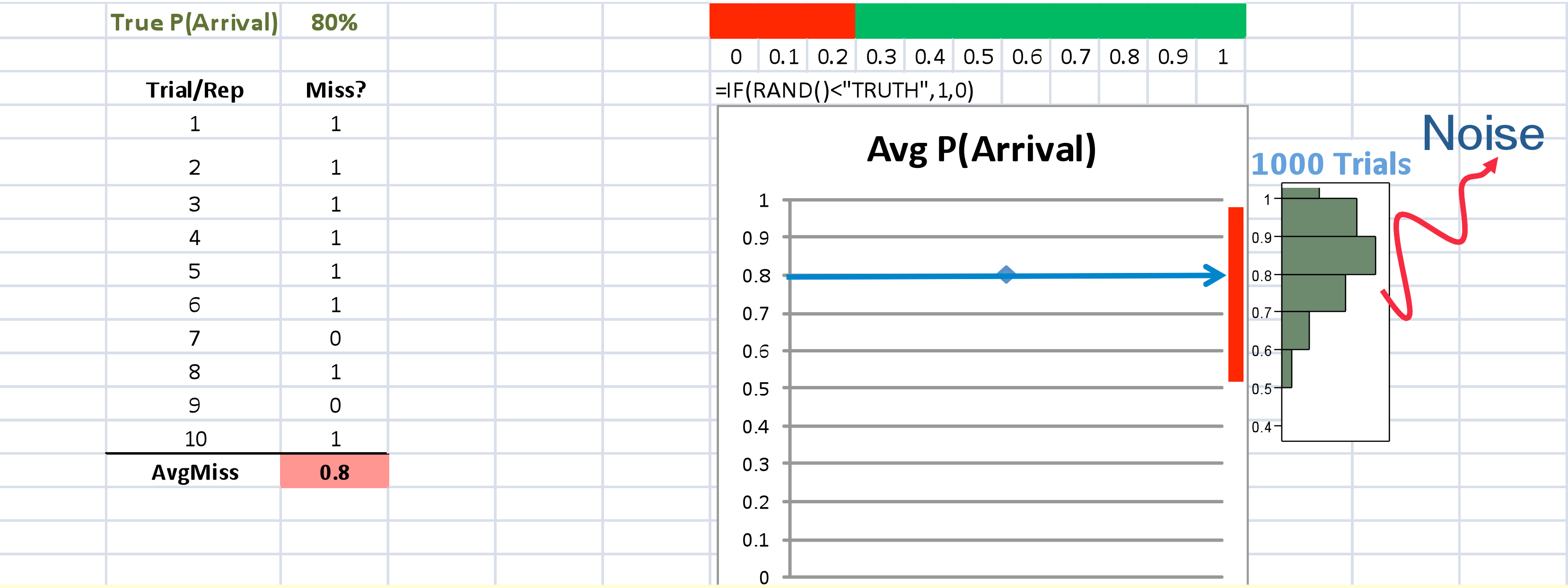
Start -- Blank Sheet of Paper: How Many?

- Let’s draw a sample of _*n*_ drops
- How many is enough to get it *right*?
  - 3 – because that’s how much $/time we have
  - 8 – because I’m an 8-guy
  - 10 – because I’m challenged by fractions
  - 30 – because something good happens at 30!

- Let’s start with 10 and see …

=> Switch to Excel File – JPADS Pancake.xls

Embedded Excel Simulation to Address "How Many?"



	True P(Arrival)	80%
	Trial/Rep	Miss?
	1	1
	2	1
	3	1
	4	1
	5	1
	6	1
	7	0
	8	1
	9	0
	10	1
	AvgMiss	0.8

0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1

```
=IF(RAND()<"TRUTH",1,0)
```

Definitions:

α - false positive error rate - concluding a difference exists (good or bad) when the difference is noise.

Confidence is 1-α.

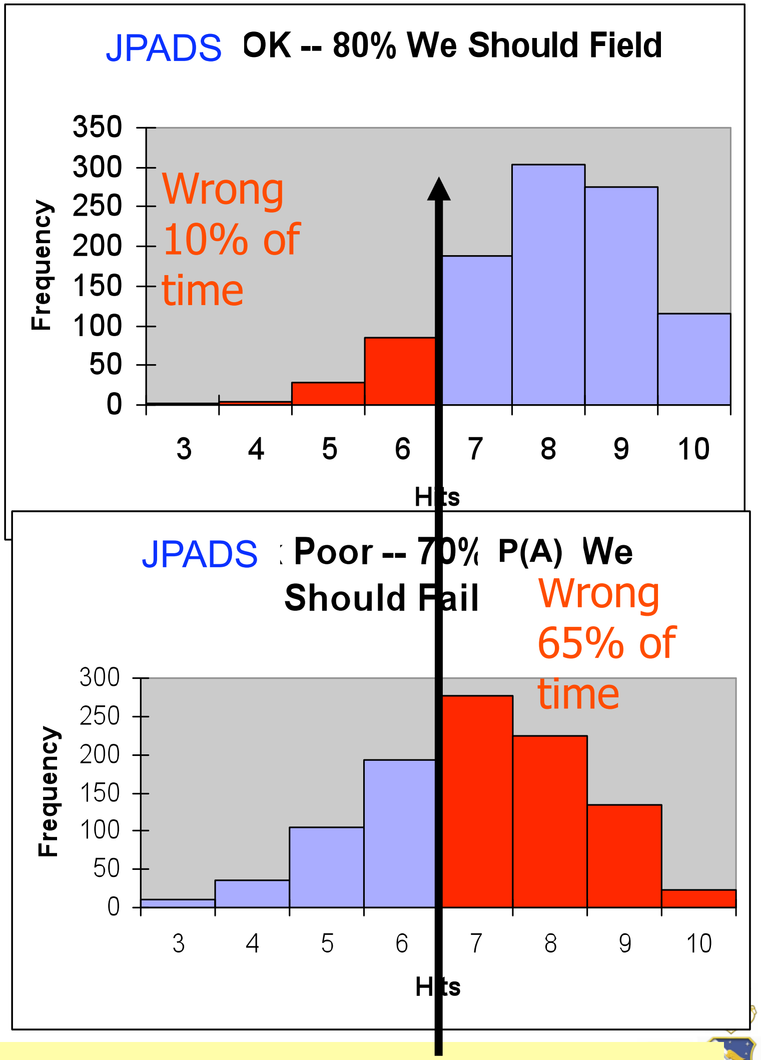
β - false negative error rate - failing to detect a difference when a difference is causally-based

Power is 1- β.

We replicate to overcome sampling error but fail to quantify the *uncertainty* in our estimates.

We seek to balance our chance of (Type I and II) errors

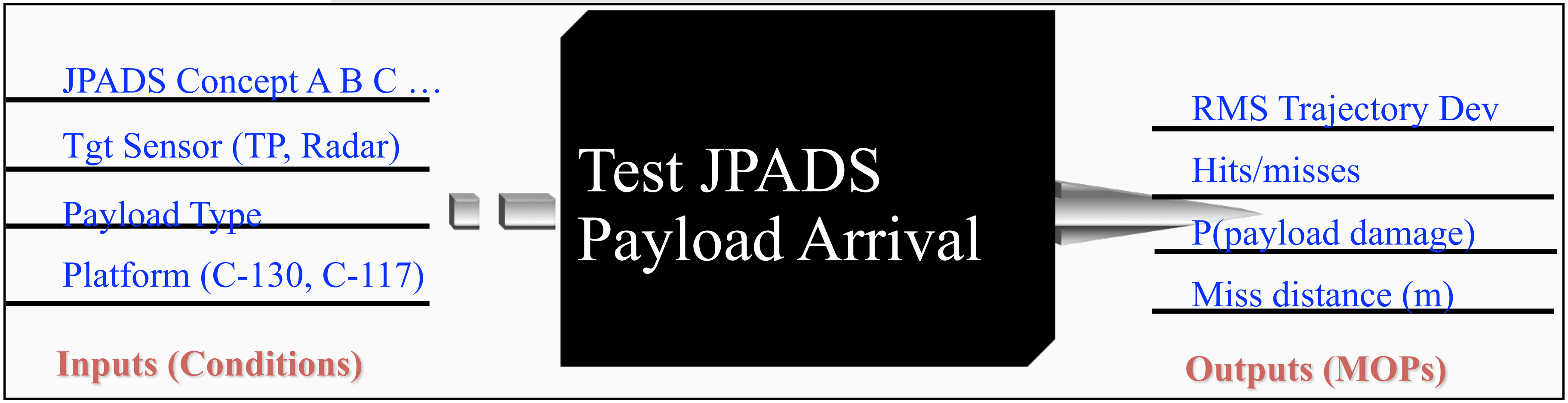
- Combining, we can trade one error for other (α for β)
- We can also increase sample size to decrease our risks in testing
- These statements not opinion –<u>mathematical fact</u> and an inescapable challenge in testing
- There are two *other* ways out … factorial designs and real-valued MOPs



Enough to Get It Right: ***Confidence*** in stating no faults; ***Power*** to detect important differences

Question 2: Which Points? How Designed Experiments Solve This

Designed Experiment *(n). Purposeful control of the inputs (factors) in such a way as to deduce their relationships (if any) with the output (responses).*



JPADS Concept A B C …

Tgt Sensor (TP, Radar)

Payload Type

Platform (C-130, C-117)

Inputs (Conditions)

Test JPADS Payload Arrival

RMS Trajectory Dev

Hits/misses

P(payload damage)

Miss distance (m)

Outputs (MOPs)

Statistician G.E.P Box said …

"All math models are false …but some are useful."

"All experiments are designed … most, poorly."

Battlespace Conditions for JPADS Case

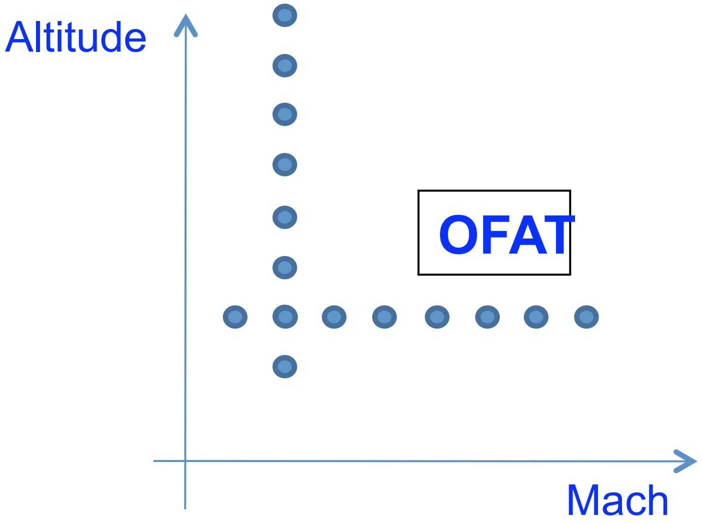
- Systems Engineering Question: Does JPADS perform at required capability level across the planned battlespace?

Type	Measure of Performance
Objective	Target acquisition range
	Target Standoff (altitude)
	launch range
	mean radial arrival distance
	probability of damage
	reliability
Subjective	Interoperability
	human factors
	tech data
	support equipment
	tactics

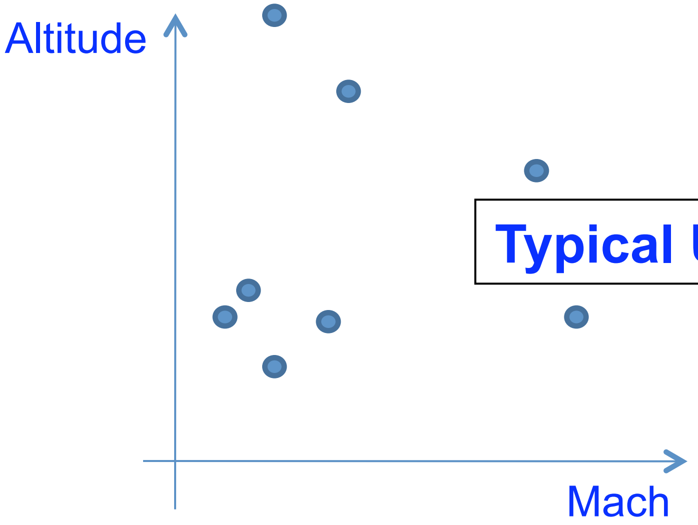
Conditions	Settings	# Levels
JPADS Variant:	A, B, C, D	4
Launch Platform:	C-130, C-17, C-5	3
Launch Opening	Ramp, Door	2
Target:	Plains, Mountain	2
Time of Day:	Dawn/Dusk, Mid-Day	3
Environment:	Forest, Desert, Snow	3
Weather:	Clear (+7nm), Haze (3-7nm), Low Ceiling/Visibility (<3000/3nm)	3
Humidity:	Low (<30%), Medium (31-79%), High (>80%)	3
Attack Azimuth:	Sun at back, Sun at beam, Sun on nose	3
Attack Altitude:	Low (<5000'), High (>5000')	2
Attack Airspeed:	Low (Mach .5), Medium (Mach .72), High (Mach .8)	3
JPADS Mode:	Autonomous, Laser Guidance	2
	Combinations	**139,968**

12 Dimensions - Obviously a large test envelope … how to search it?

Spanning the Battlespace – Traditional Test Designs

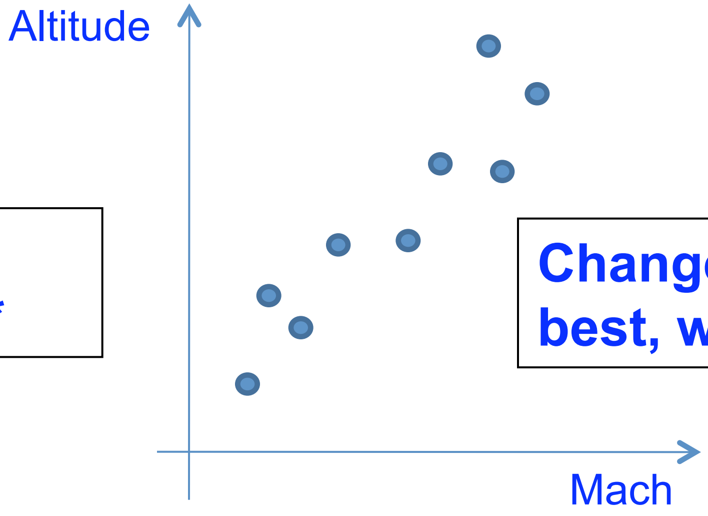


OFAT



Typical Use Cases

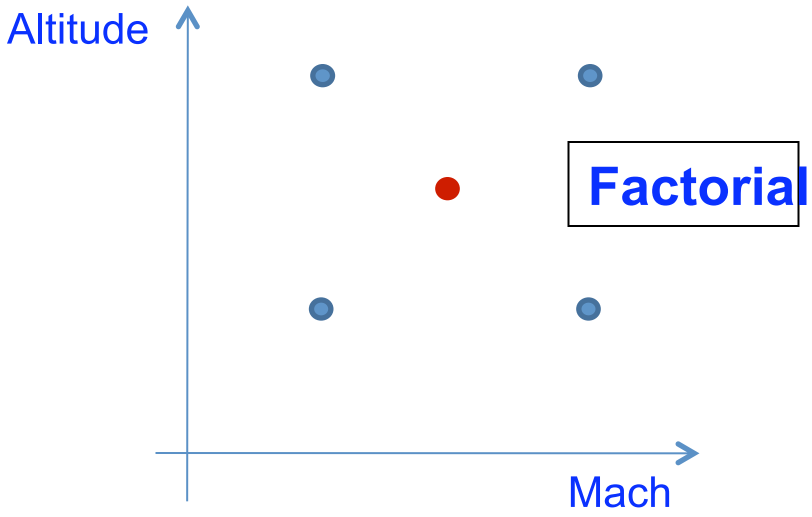
And … the always popular DWWDLT*

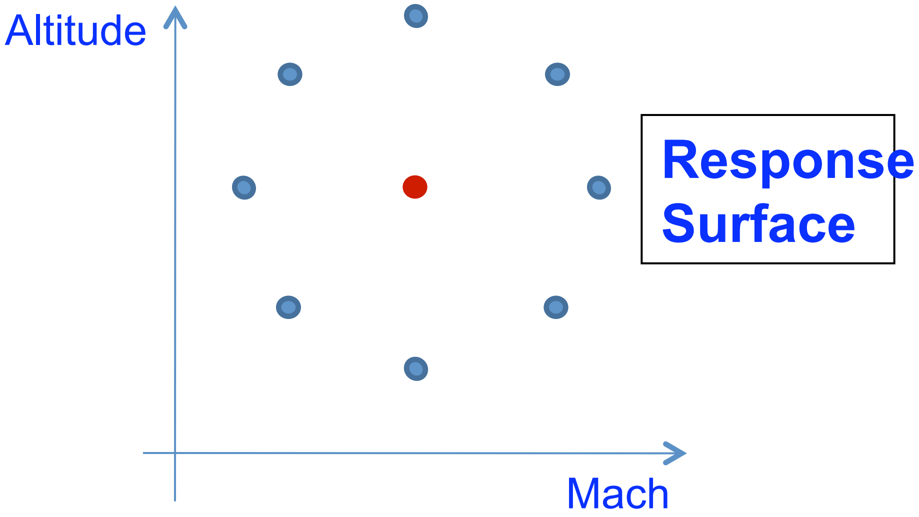


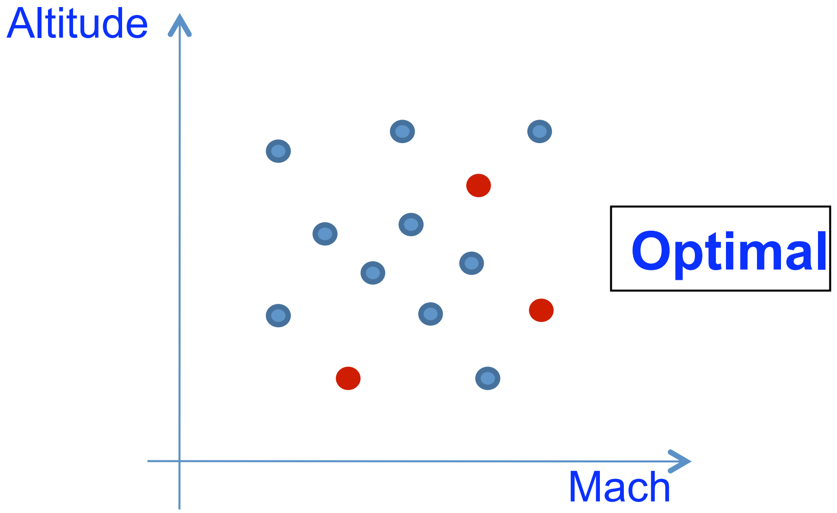
Change variables together: best, worst, nominal

* Do What We Did Last Time

Spanning the Battlespace - DOE



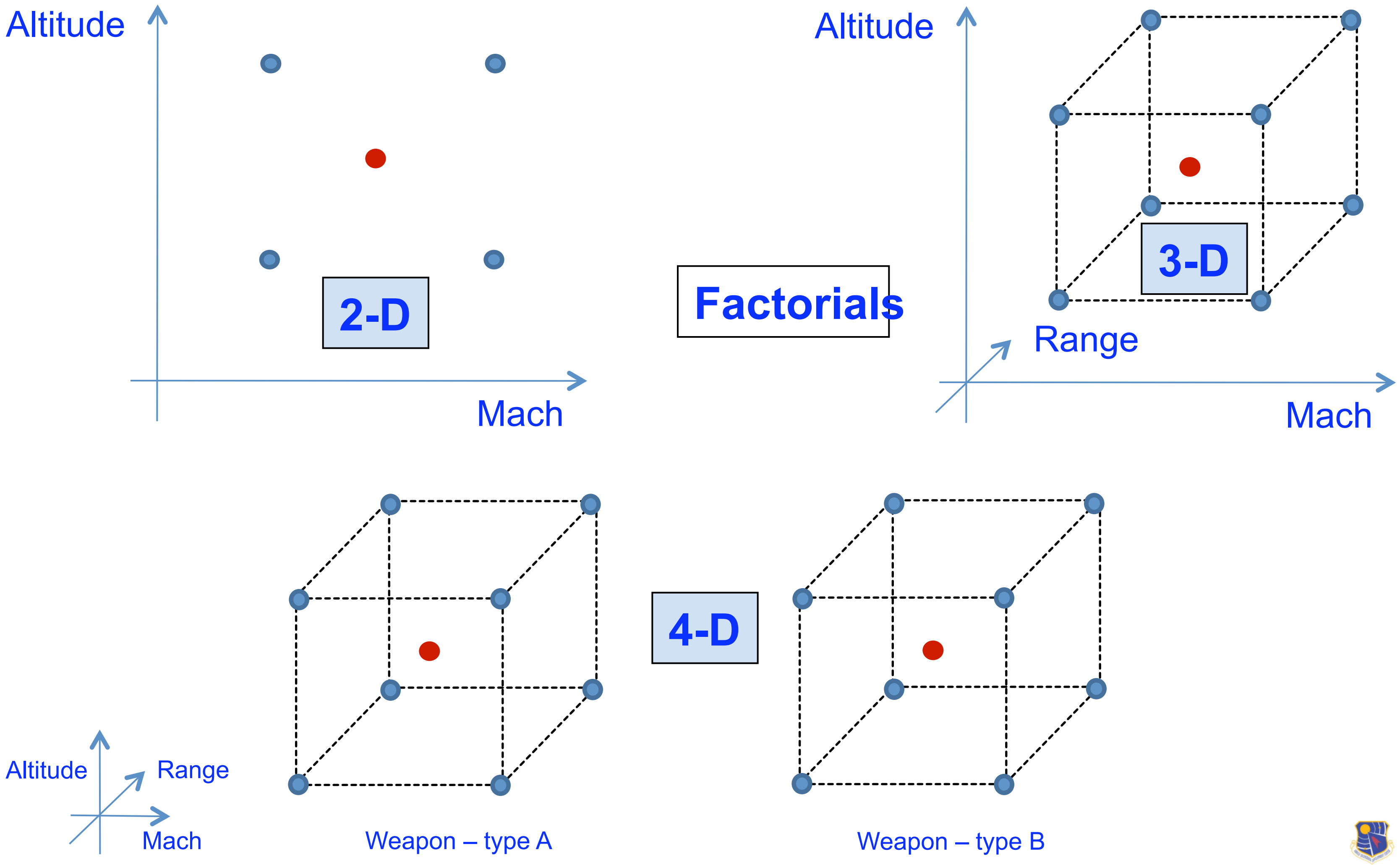




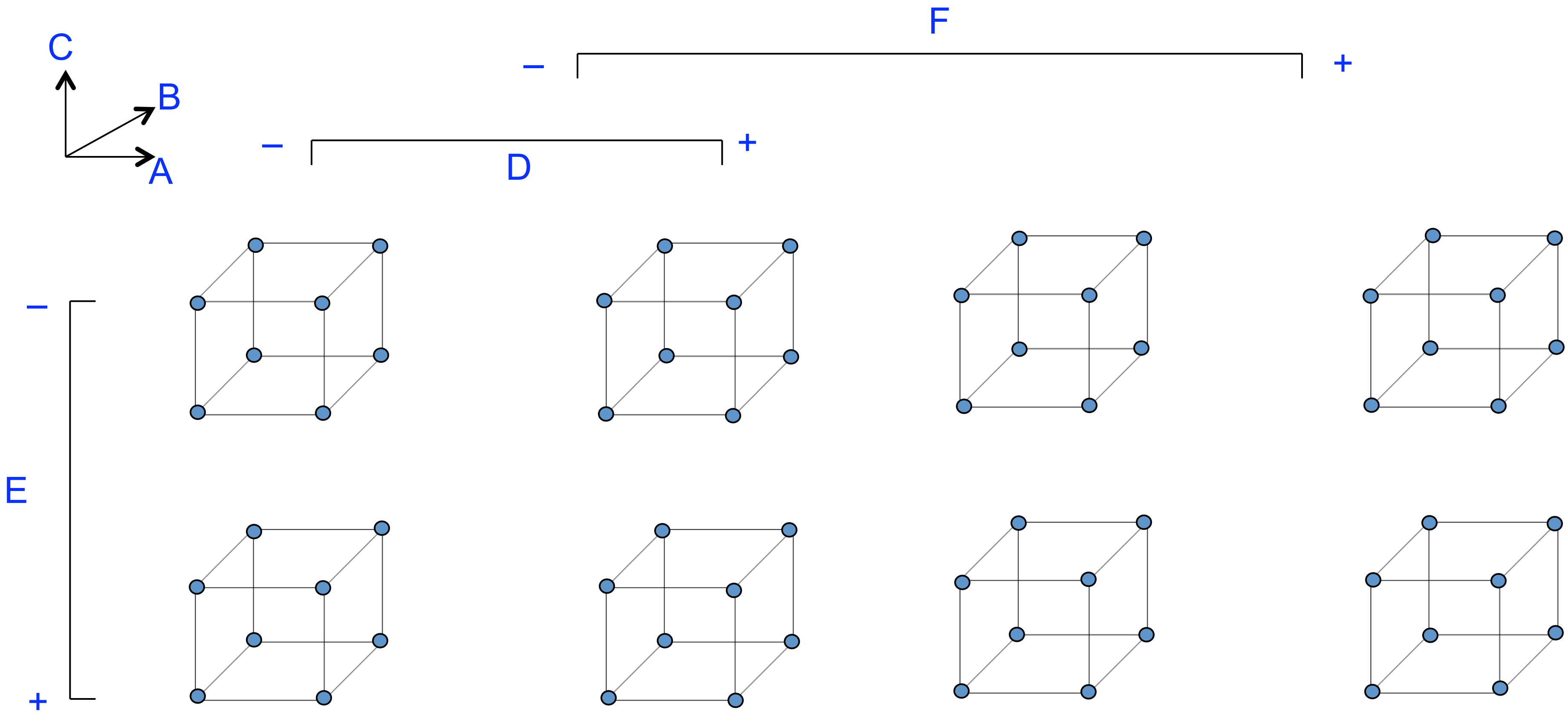
● single point

● replicate

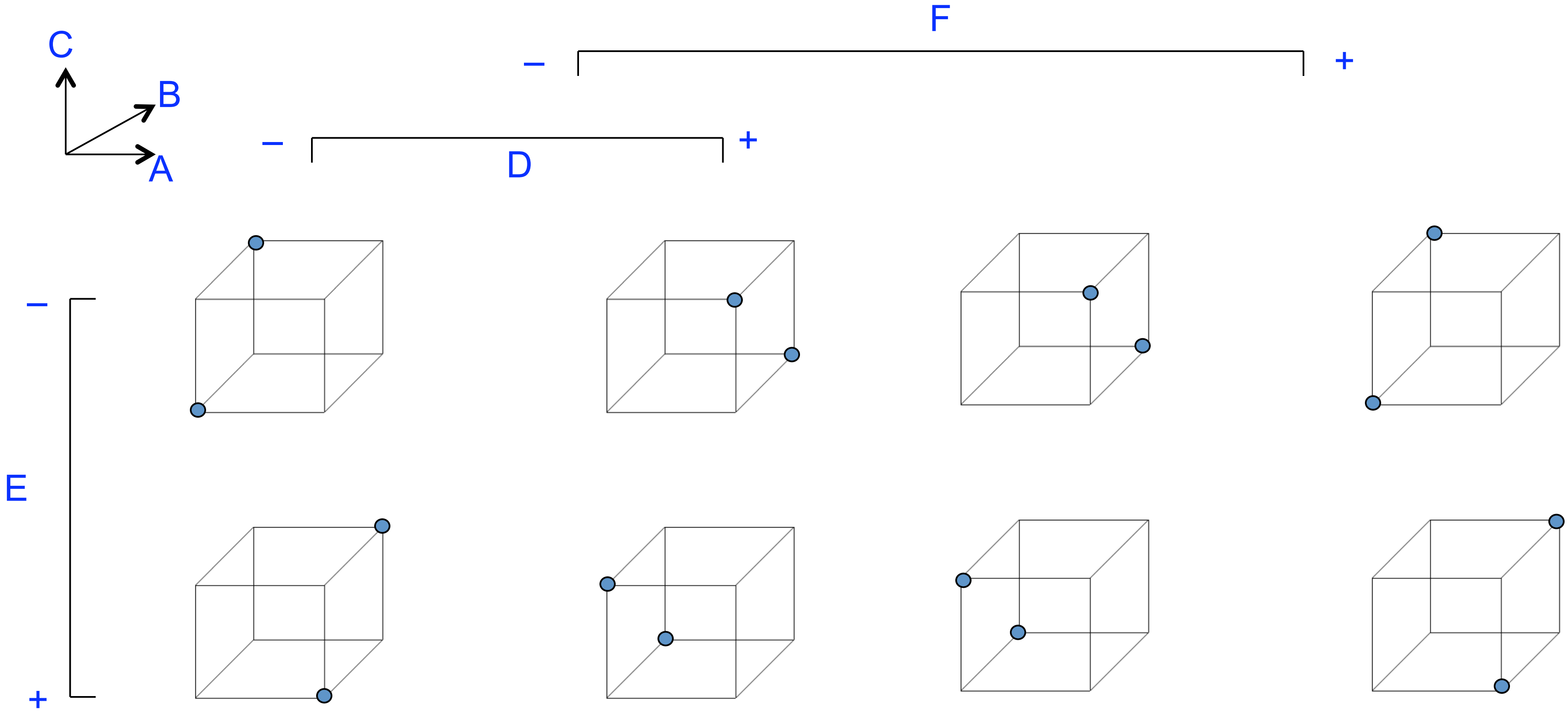
More Variables – DOE Factorials



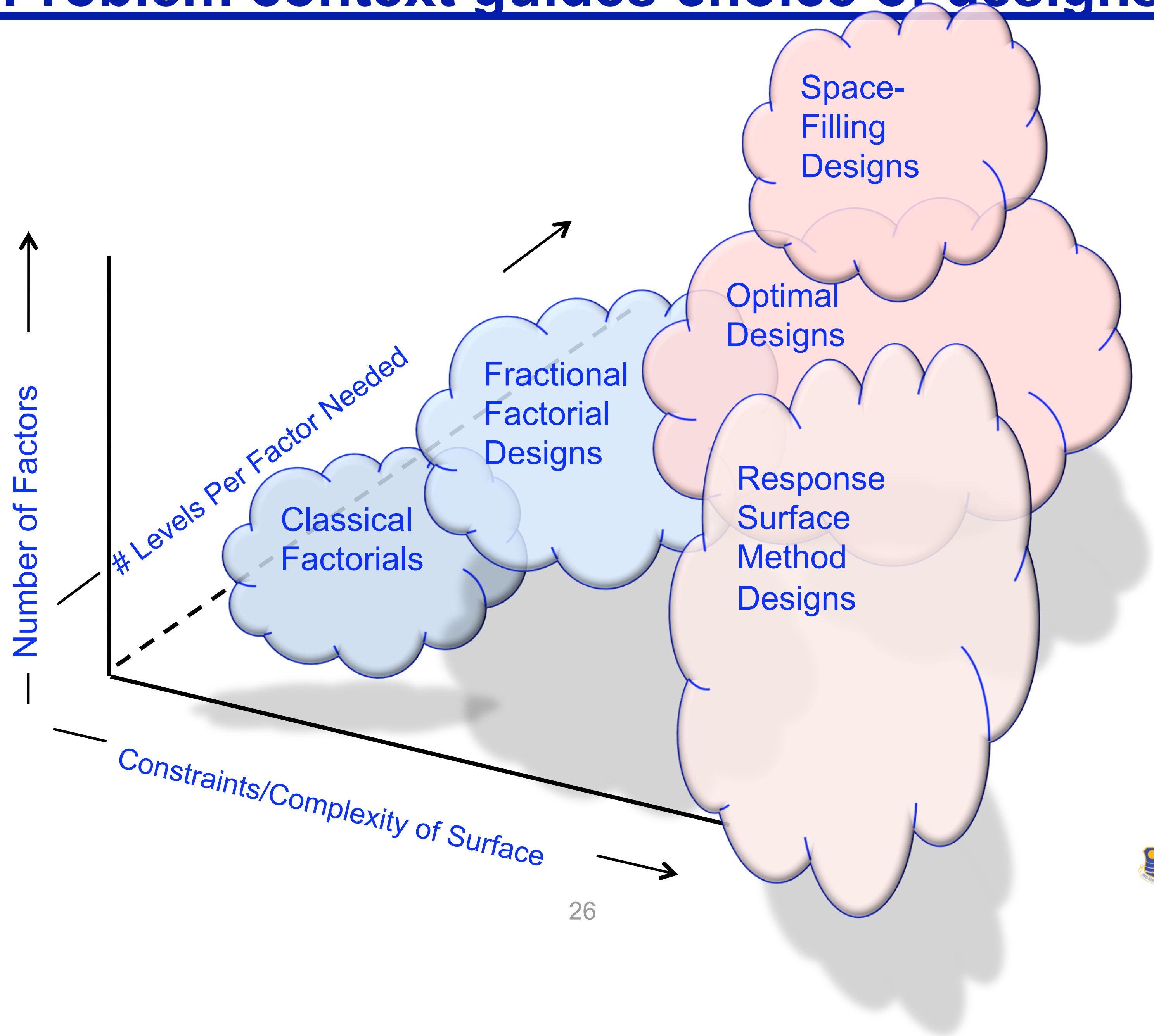
Even More Variables (here – 6)



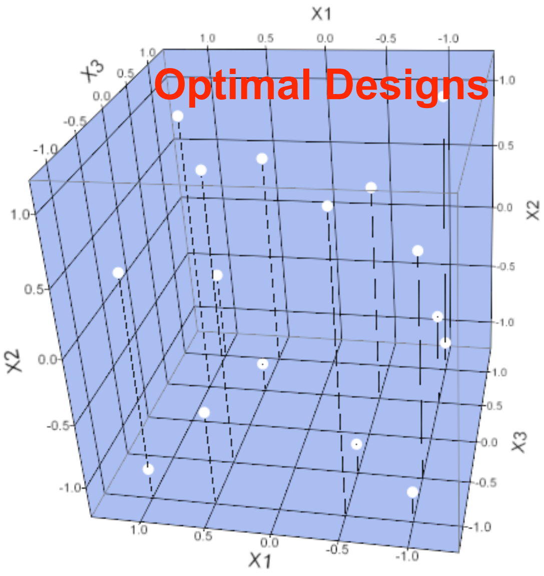
Efficiencies in Test - Fractions

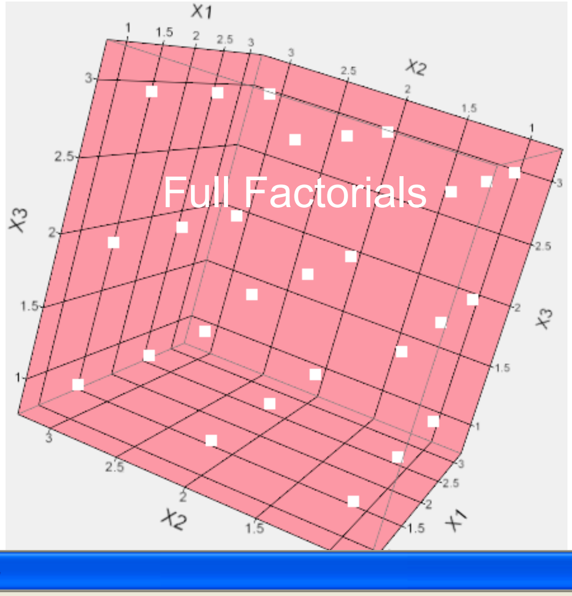


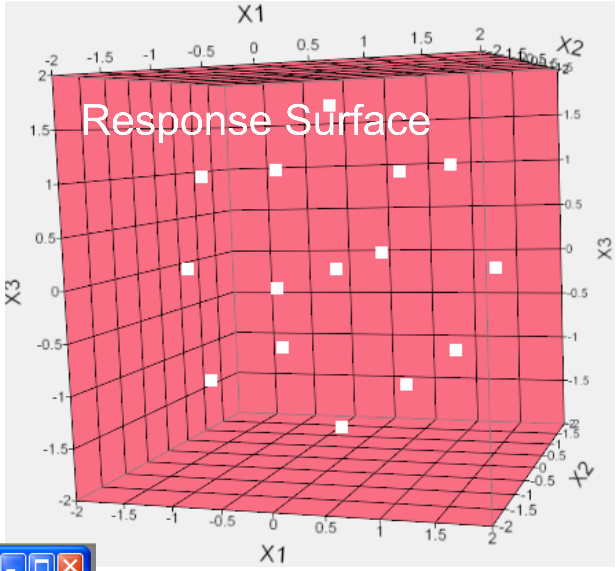
Problem context guides choice of designs

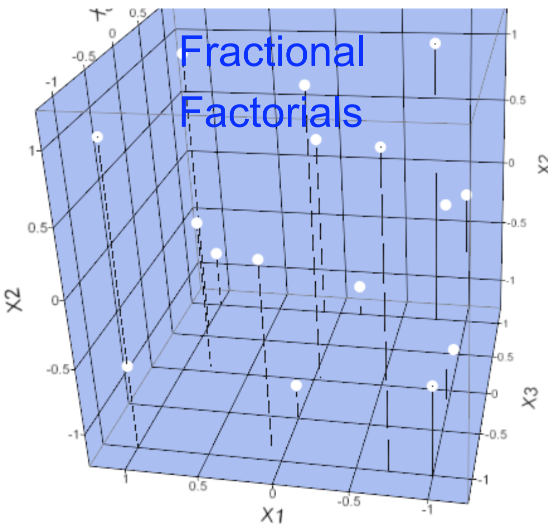

Space-Filling Designs
Optimal Designs
Fractional Factorial Designs
Classical Factorials
Response Surface Method Designs
Number of Factors
# Levels Per Factor Needed
Constraints/Complexity of Surface

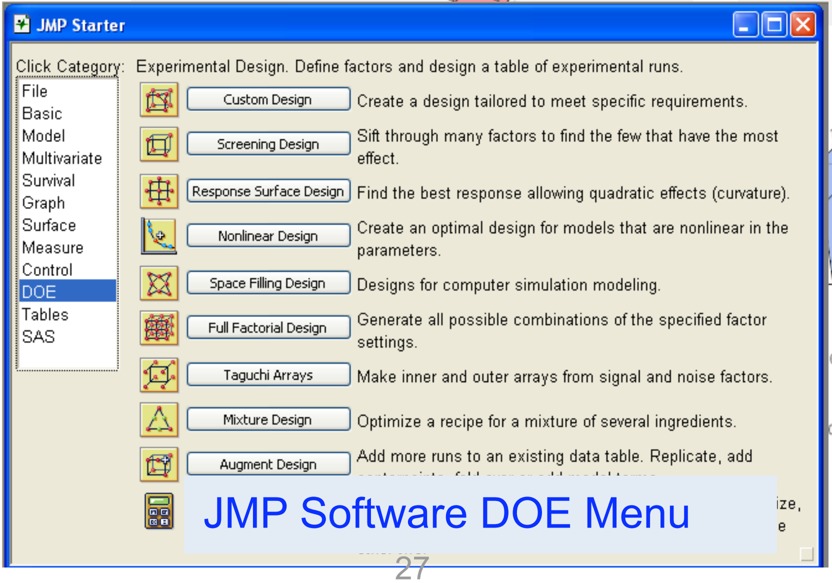
We have a wide menu of design choices with DOE







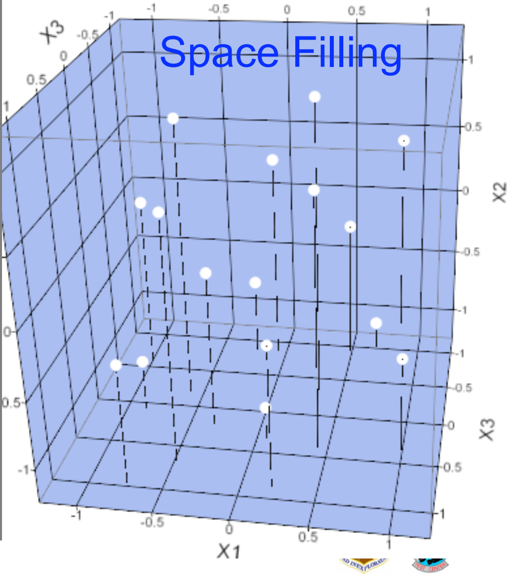




Experimental Design. Define factors and design a table of experimental runs.

Custom Design	Create a design tailored to meet specific requirements.
Screening Design	Sift through many factors to find the few that have the most effect.
Response Surface Design	Find the best response allowing quadratic effects (curvature).
Nonlinear Design	Create an optimal design for models that are nonlinear in the parameters.
Space Filling Design	Designs for computer simulation modeling.
Full Factorial Design	Generate all possible combinations of the specified factor settings.
Taguchi Arrays	Make inner and outer arrays from signal and noise factors.
Mixture Design	Optimize a recipe for a mixture of several ingredients.
Augment Design	Add more runs to an existing data table. Replicate, add
	ize, e

JMP Software DOE Menu



Which Points to Span the Relevant Battlespace?

4 reps 1 var

JPADS A	JPADS B
4	4

- Factorial (crossed) designs let us *learn more* from the same number of assets
- We can also use Factorials to *reduce assets* while maintaining confidence and power
- Or we can *combine* the two

All four Designs share the same ***power*** and ***confidence***

- How to support such an amazing claim?

2 reps 2 vars

	JPADS A	JPADS B
Ammo	2	2
Food	2	2

1 reps 3 vars

		JPADS A	JPADS B
Eglin (Low)	Ammo	1	1
	Food	1	1
Nellis (High)	Ammo	1	1
	Food	1	1

½ rep 4 vars

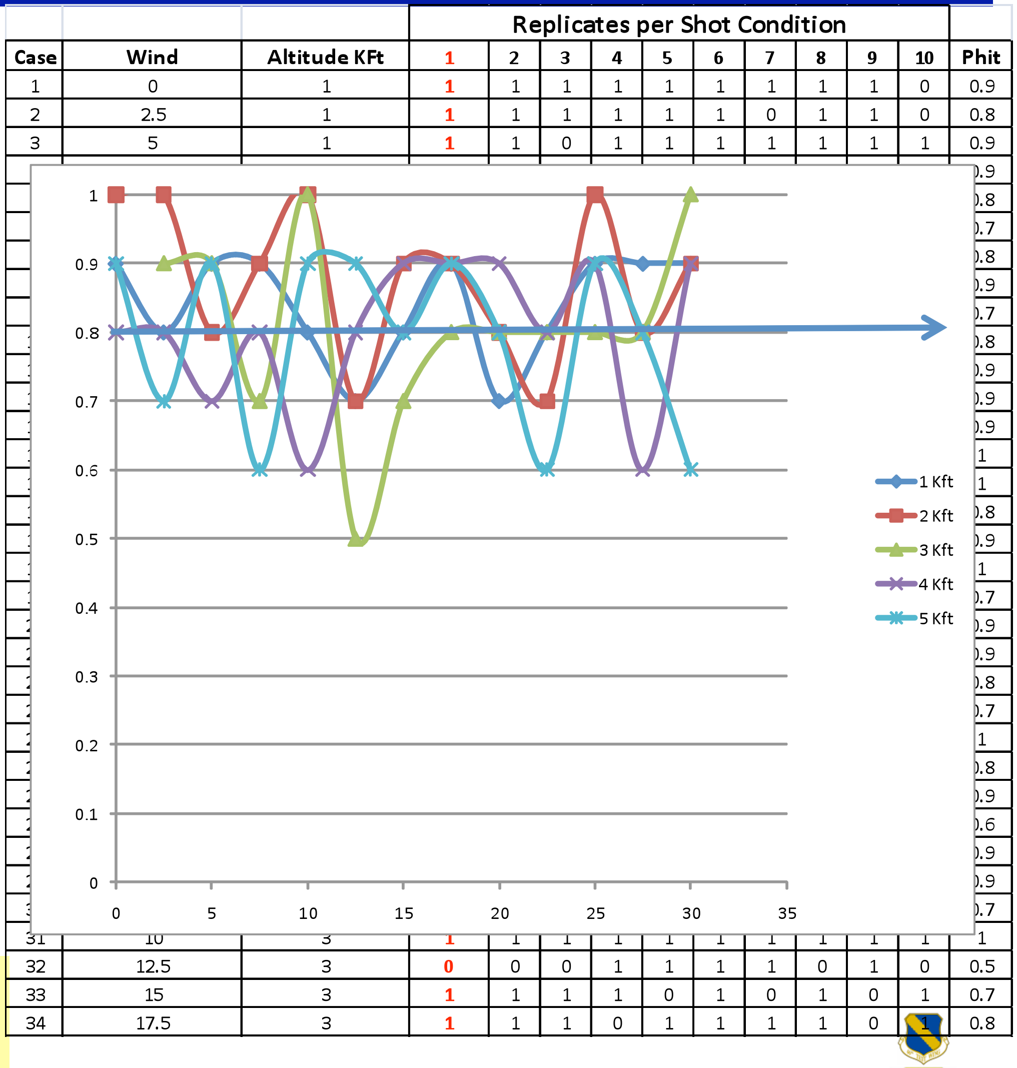
			JPADS A	JPADS B
Dawn (low light)	Eglin (Low)	Ammo	1	
		Food		1
	Nellis (High)	Ammo		1
		Food	1	
Midday (bright)	Eglin (Low)	Ammo		1
		Food	1	
	Nellis (High)	Ammo	1	
		Food		1

=> Switch to Excel File – JPADS Pancake.xls

Equal Power? A preposterous claim … how to justify it?

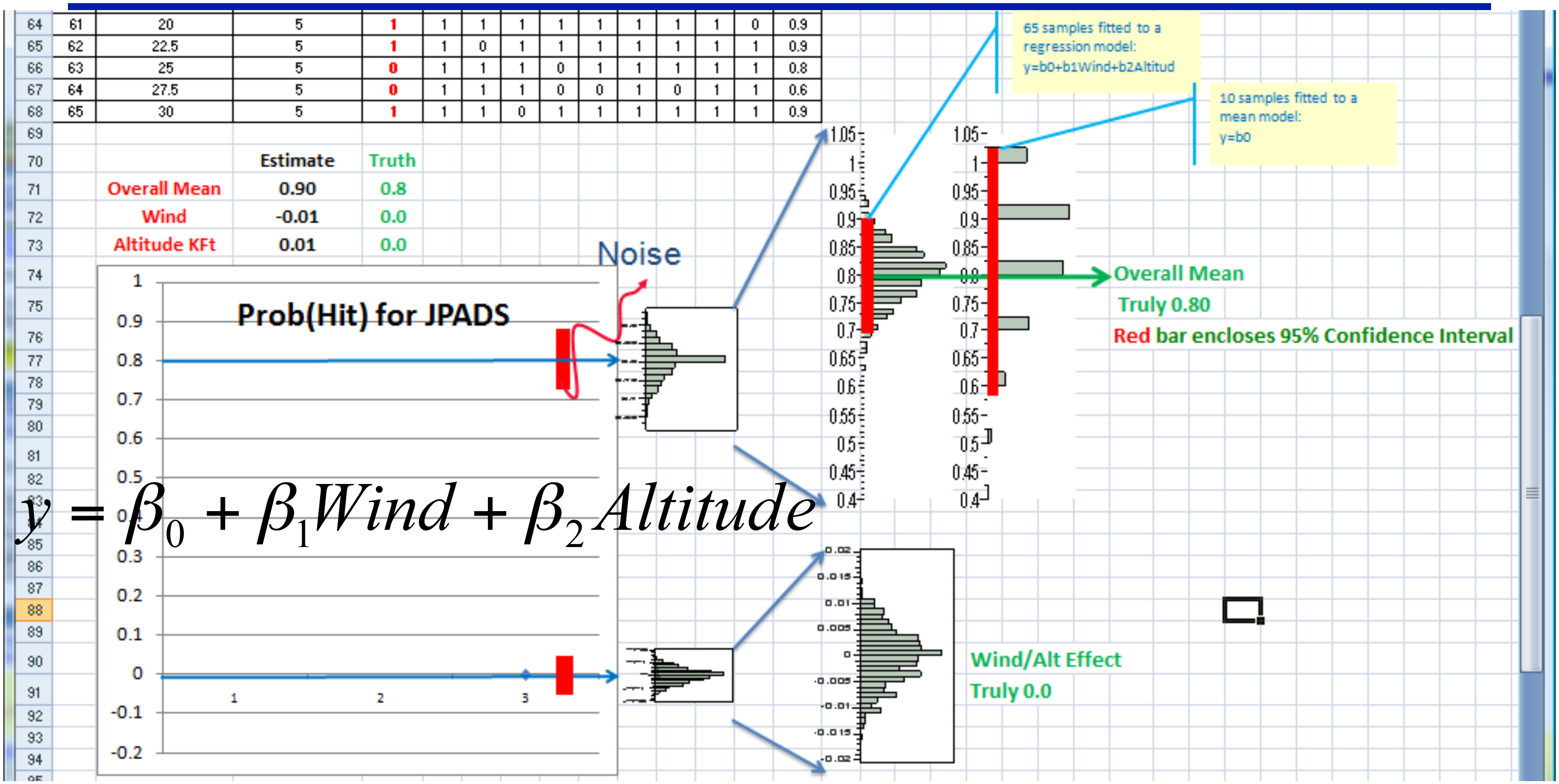
- Consider again our JPADS problem across 2 dimensions
- 13 wind speeds x 5 altitudes = 65 cases x 10 reps each = 650 trials
- Surely this will solve our problem with noise?

It will **not** … we have 65 separate 10-sample trials



			Replicates per Shot Condition										
Case	Wind	Altitude KFt	1	2	3	4	5	6	7	8	9	10	Phit
1	0	1	1	1	1	1	1	1	1	1	1	0	0.9
2	2.5	1	1	1	1	1	1	1	0	1	1	0	0.8
3	5	1	1	1	0	1	1	1	1	1	1	1	0.9
													.9
													.8
													.7
													.8
													.9
													.7
													.8
													.9
													.9
													.9
													1
													1
													.8
													.9
													1
													.7
													.9
													.9
													.8
													.7
													1
													.8
													.9
													.6
													.9
													.9
													.7
31	10	3	1	1	1	1	1	1	1	1	1	1	1
32	12.5	3	0	0	0	1	1	1	1	0	1	0	0.5
33	15	3	1	1	1	1	0	1	0	1	0	1	0.7
34	17.5	3	1	1	1	0	1	1	1	1	0	1	0.8

But, discard 9/10th of trials … strap 1/10th into a math model



64	61	20	5	1	1	1	1	1	1	1	1	1	0	0.9
65	62	22.5	5	1	1	0	1	1	1	1	1	1	1	0.9
66	63	25	5	0	1	1	1	0	1	1	1	1	1	0.8
67	64	27.5	5	0	1	1	1	0	0	1	0	1	1	0.6
68	65	30	5	1	1	1	0	1	1	1	1	1	1	0.9

	Estimate	Truth
Overall Mean	0.90	0.8
Wind	-0.01	0.0
Altitude KFt	0.01	0.0

$$y = \beta_0 + \beta_1 Wind + \beta_2 Altitude$$

65 samples fitted to a regression model: y=b0+b1Wind+b2Altitud

10 samples fitted to a mean model: y=b0

Overall Mean Truly 0.80

Red bar encloses 95% Confidence Interval

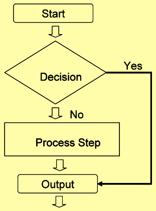
Wind/Alt Effect Truly 0.0

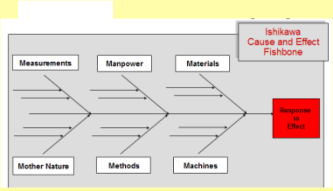
DOE math model straps all the physics together:
- *reducing* samples per condition by 90% while
- *increasing* our prediction accuracy 50%

Note: this speaks to the method of analysis (Challenge #4.)

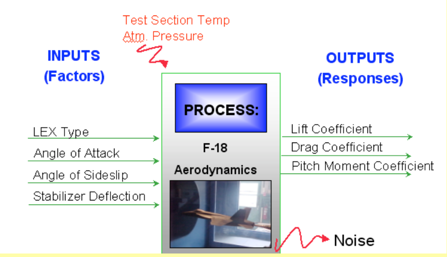
Test as Science vs. Art: Experimental Design Test Process is Well-Defined

Planning: Factors
Desirable and Nuisance

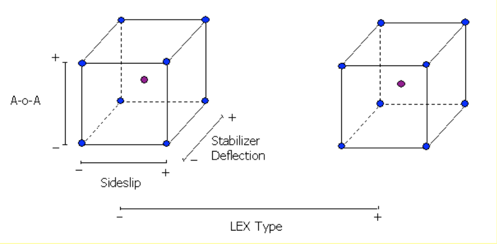




Desired Factors
and Responses

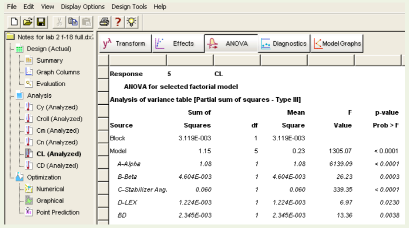


Design Points



Test Matrix

Randomize & Block ->
Results and Analysis



Response 5 CL

ANOVA for selected factorial model

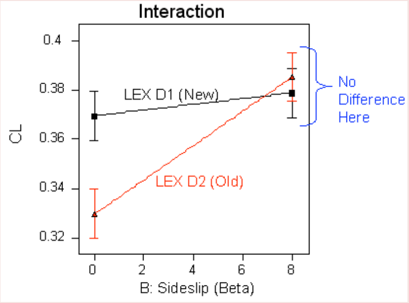
Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F
Block	3.119E-003	1	3.119E-003		
Model	1.15	5	0.23	1305.07	< 0.0001
A-Alpha	*1.08*	*1*	*1.08*	*6139.09*	*< 0.0001*
B-Beta	*4.604E-003*	*1*	*4.604E-003*	*26.23*	*0.0003*
C-Stabilizer Ang.	*0.060*	*1*	*0.060*	*339.35*	*< 0.0001*
D-LEX	*1.224E-003*	*1*	*1.224E-003*	*6.97*	*0.0230*
BD	*2.345E-003*	*1*	*2.345E-003*	*13.36*	*0.0038*

Model Build

C_L = +0.38
+0.26 x A-o-A
+0.017 x Sideslip
+0.061 x Stabilizer Deflection
-.00875 x LEX Type
+0.012 x Sideslip x LEX Type

Discovery, Understanding
Prediction, Re-design

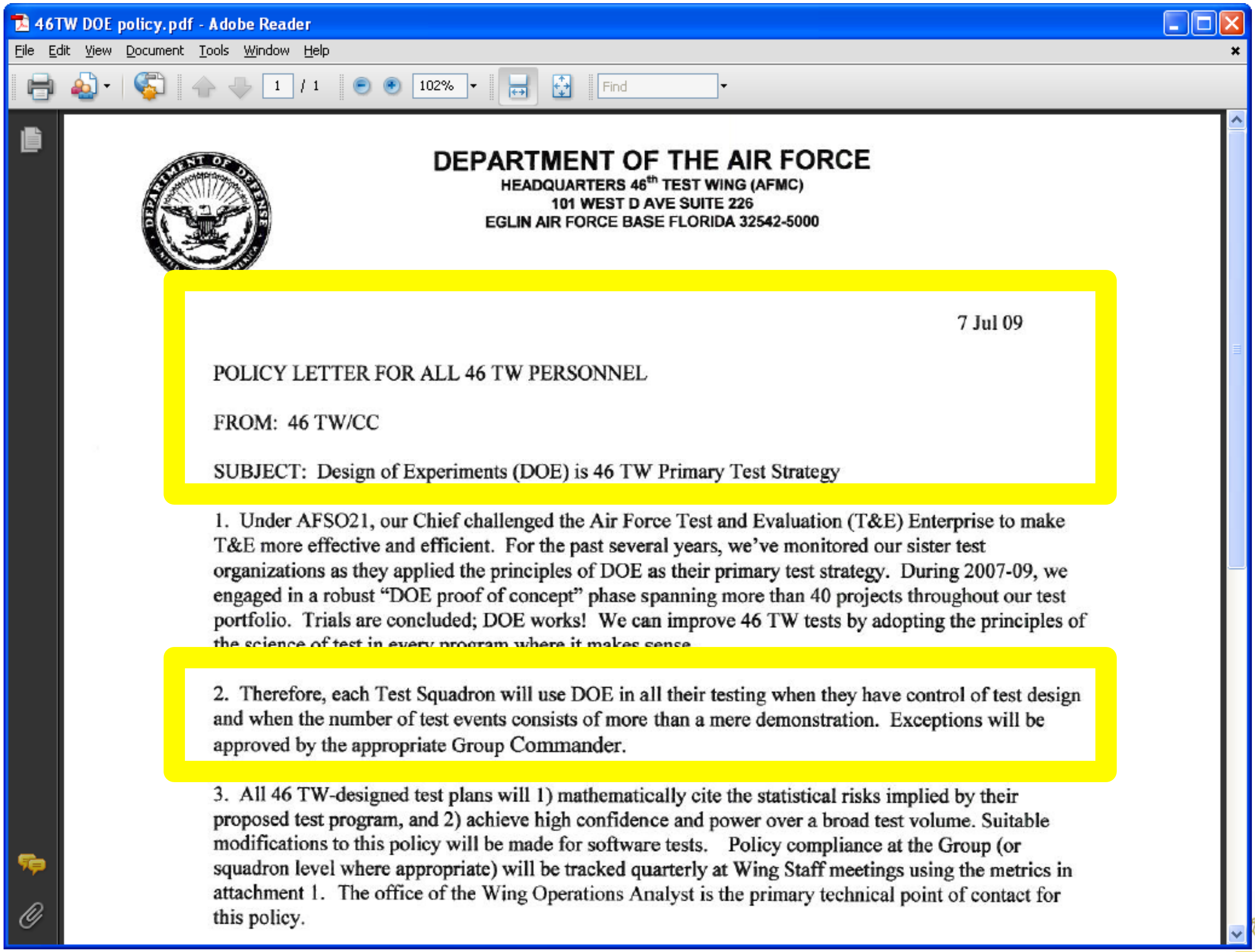


It applies to our tests: DOE in 50+ operations over 20 years

- IR Sensor Predictions
- Ballistics 6 DOF Initial Conditions
- Wind Tunnel fuze characteristics
- Camouflaged Target JT&E ($30M)
- AC-130 40/105mm gunfire CEP evals
- AMRAAM HWIL test facility validation
- 60+ ECM development + RWR tests
- GWEF Maverick sensor upgrades
- 30mm Ammo over-age LAT testing
- Contact lens plastic injection molding
- 30mm gun DU/HEI accuracy (A-10C)
- GWEF ManPad Hit-point prediction
- AIM-9X Simulation Validation
- Link 16 and VHF/UHF/HF Comm tests
- TF radar flight control system gain opt
- New FCS software to cut C-17 PIO
- AIM-9X+JHMCS Tactics Development
- MAU 169/209 LGB fly-off and eval

- Characterizing Seek Eagle Ejector Racks
- SFW altimeter false alarm trouble-shoot
- TMD safety lanyard flight envelope
- Penetrator & reactive frag design
- F-15C/F-15E Suite 4 + Suite 5 OFPs
- PLAID Performance Characterization
- JDAM, LGB weapons accuracy testing
- Best Autonomous seeker algorithm
- SAM Validation versus Flight Test
- ECM development ground mounts (10’s)
- AGM-130 Improved Data Link HF Test
- TPS A-G WiFi characterization
- MC/EC-130 flare decoy characterization
- SAM simulation validation vs. live-fly
- Targeting Pod TLE estimates
- Chem CCA process characterization
- Medical Oxy Concentration T&E
- Multi-MDS Link 16 and Rover video test

Adopt a Policy of Well-Designed Tests


46TW DOE policy.pdf - Adobe Reader
File Edit View Document Tools Window Help
1 / 1
102%
Find

DEPARTMENT OF THE AIR FORCE
HEADQUARTERS 46th TEST WING (AFMC)
101 WEST D AVE SUITE 226
EGLIN AIR FORCE BASE FLORIDA 32542-5000

7 Jul 09

POLICY LETTER FOR ALL 46 TW PERSONNEL

FROM: 46 TW/CC

SUBJECT: Design of Experiments (DOE) is 46 TW Primary Test Strategy

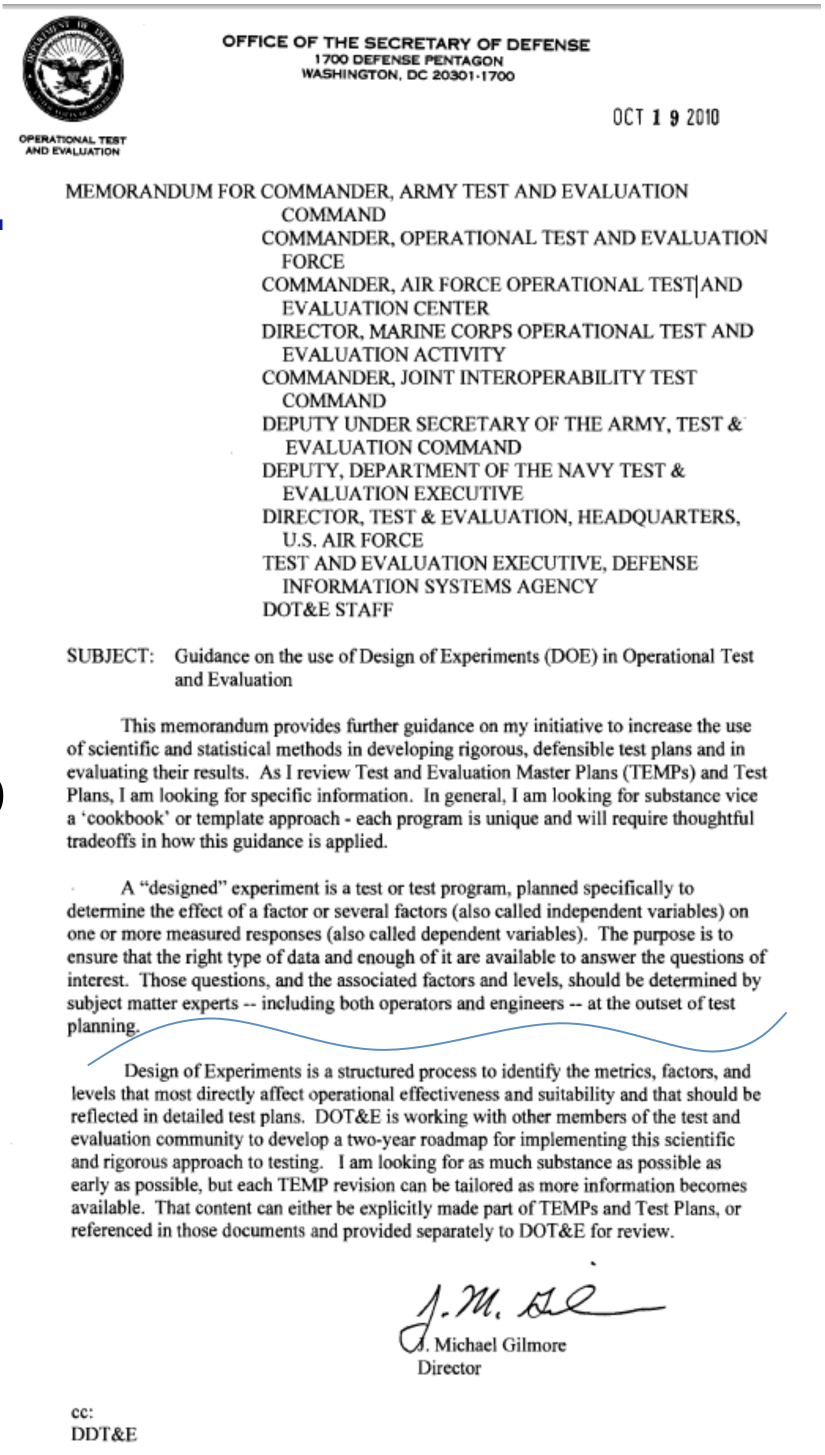
1. Under AFSO21, our Chief challenged the Air Force Test and Evaluation (T&E) Enterprise to make T&E more effective and efficient. For the past several years, we've monitored our sister test organizations as they applied the principles of DOE as their primary test strategy. During 2007-09, we engaged in a robust "DOE proof of concept" phase spanning more than 40 projects throughout our test portfolio. Trials are concluded; DOE works! We can improve 46 TW tests by adopting the principles of the science of test in every program where it makes sense.

2. Therefore, each Test Squadron will use DOE in all their testing when they have control of test design and when the number of test events consists of more than a mere demonstration. Exceptions will be approved by the appropriate Group Commander.

3. All 46 TW-designed test plans will 1) mathematically cite the statistical risks implied by their proposed test program, and 2) achieve high confidence and power over a broad test volume. Suitable modifications to this policy will be made for software tests. Policy compliance at the Group (or squadron level where appropriate) will be tracked quarterly at Wing Staff meetings using the metrics in attachment 1. The office of the Wing Operations Analyst is the primary technical point of contact for this policy.

Checklist: Fruits of Well-Designed Tests

- □Specify Goal/Objective
- □List Quantitative Responses
- □List factors/levels & how to control in test
- □Strategy to place Points
- □Compute Confidence/ Power



OFFICE OF THE SECRETARY OF DEFENSE
1700 DEFENSE PENTAGON
WASHINGTON, DC 20301-1700

OPERATIONAL TEST
AND EVALUATION

OCT 1 9 2010

MEMORANDUM FOR COMMANDER, ARMY TEST AND EVALUATION COMMAND
COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE
COMMANDER, AIR FORCE OPERATIONAL TEST AND EVALUATION CENTER
DIRECTOR, MARINE CORPS OPERATIONAL TEST AND EVALUATION ACTIVITY
COMMANDER, JOINT INTEROPERABILITY TEST COMMAND
DEPUTY UNDER SECRETARY OF THE ARMY, TEST & EVALUATION COMMAND
DEPUTY, DEPARTMENT OF THE NAVY TEST & EVALUATION EXECUTIVE
DIRECTOR, TEST & EVALUATION, HEADQUARTERS, U.S. AIR FORCE
TEST AND EVALUATION EXECUTIVE, DEFENSE INFORMATION SYSTEMS AGENCY
DOT&E STAFF

SUBJECT: Guidance on the use of Design of Experiments (DOE) in Operational Test and Evaluation

This memorandum provides further guidance on my initiative to increase the use of scientific and statistical methods in developing rigorous, defensible test plans and in evaluating their results. As I review Test and Evaluation Master Plans (TEMPs) and Test Plans, I am looking for specific information. In general, I am looking for substance vice a 'cookbook' or template approach - each program is unique and will require thoughtful tradeoffs in how this guidance is applied.

A "designed" experiment is a test or test program, planned specifically to determine the effect of a factor or several factors (also called independent variables) on one or more measured responses (also called dependent variables). The purpose is to ensure that the right type of data and enough of it are available to answer the questions of interest. Those questions, and the associated factors and levels, should be determined by subject matter experts -- including both operators and engineers -- at the outset of test planning.

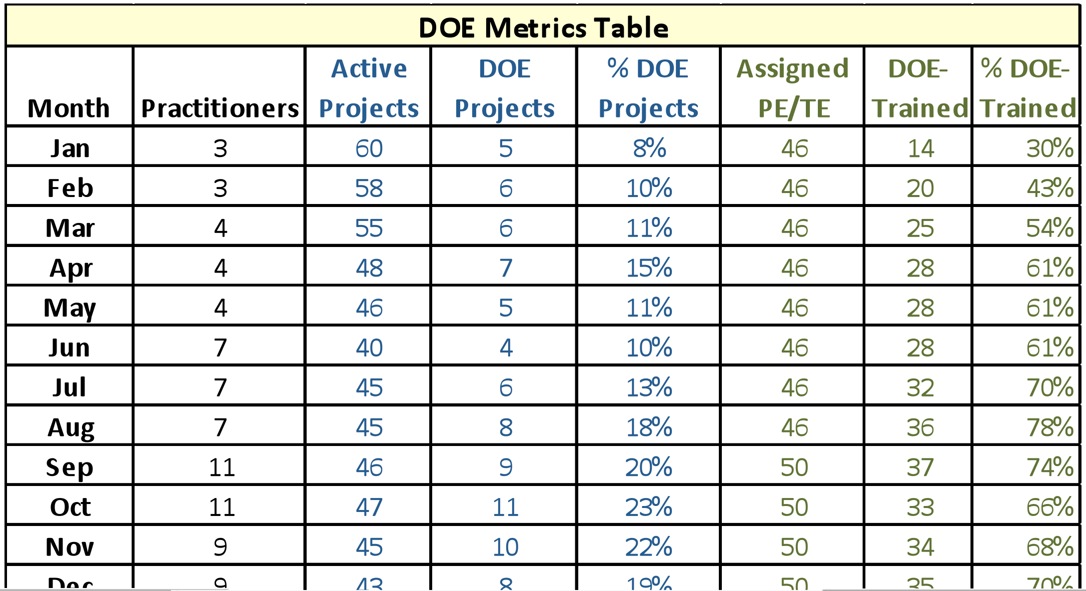
Design of Experiments is a structured process to identify the metrics, factors, and levels that most directly affect operational effectiveness and suitability and that should be reflected in detailed test plans. DOT&E is working with other members of the test and evaluation community to develop a two-year roadmap for implementing this scientific and rigorous approach to testing. I am looking for as much substance as possible as early as possible, but each TEMP revision can be tailored as more information becomes available. That content can either be explicitly made part of TEMPs and Test Plans, or referenced in those documents and provided separately to DOT&E for review.

J. M. Gil

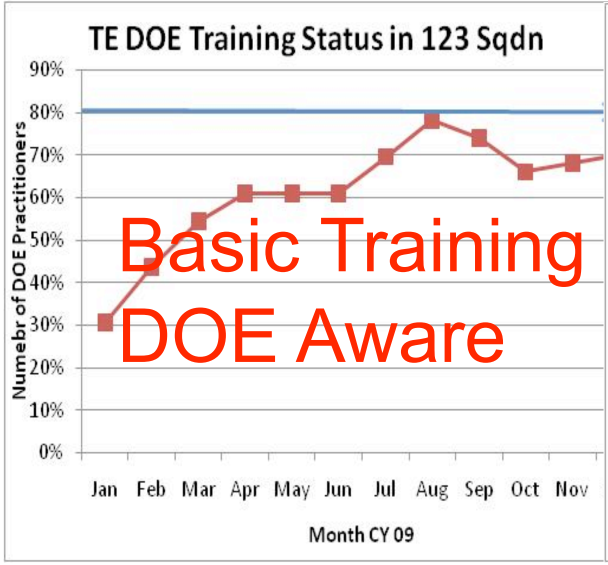
J. Michael Gilmore
Director

cc:
DDT&E

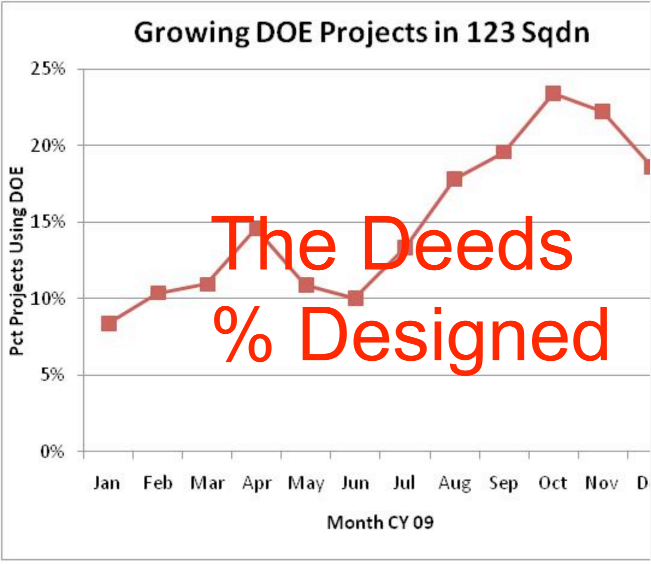
What you measure gets done …
Sample Unit Quarterly Metrics



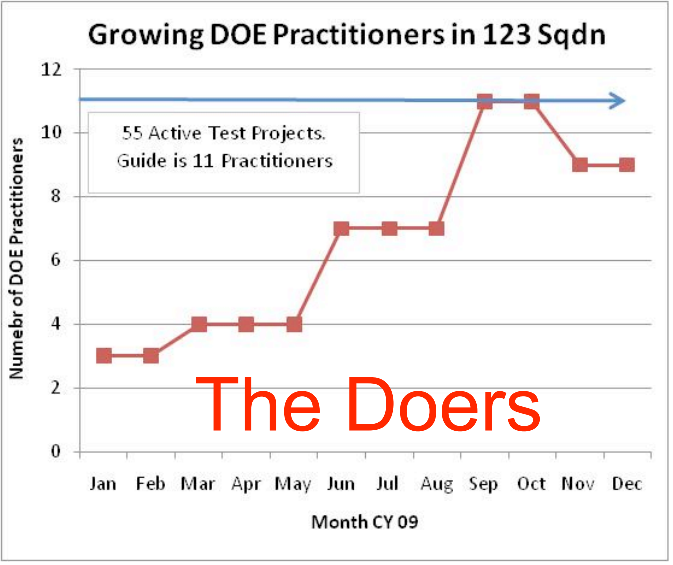
DOE Metrics Table							
Month	Practitioners	Active Projects	DOE Projects	% DOE Projects	Assigned PE/TE	DOE-Trained	% DOE-Trained
Jan	3	60	5	8%	46	14	30%
Feb	3	58	6	10%	46	20	43%
Mar	4	55	6	11%	46	25	54%
Apr	4	48	7	15%	46	28	61%
May	4	46	5	11%	46	28	61%
Jun	7	40	4	10%	46	28	61%
Jul	7	45	6	13%	46	32	70%
Aug	7	45	8	18%	46	36	78%
Sep	11	46	9	20%	50	37	74%
Oct	11	47	11	23%	50	33	66%
Nov	9	45	10	22%	50	34	68%
Dec	9	43	8	19%	50	35	70%



Basic Training
DOE Aware

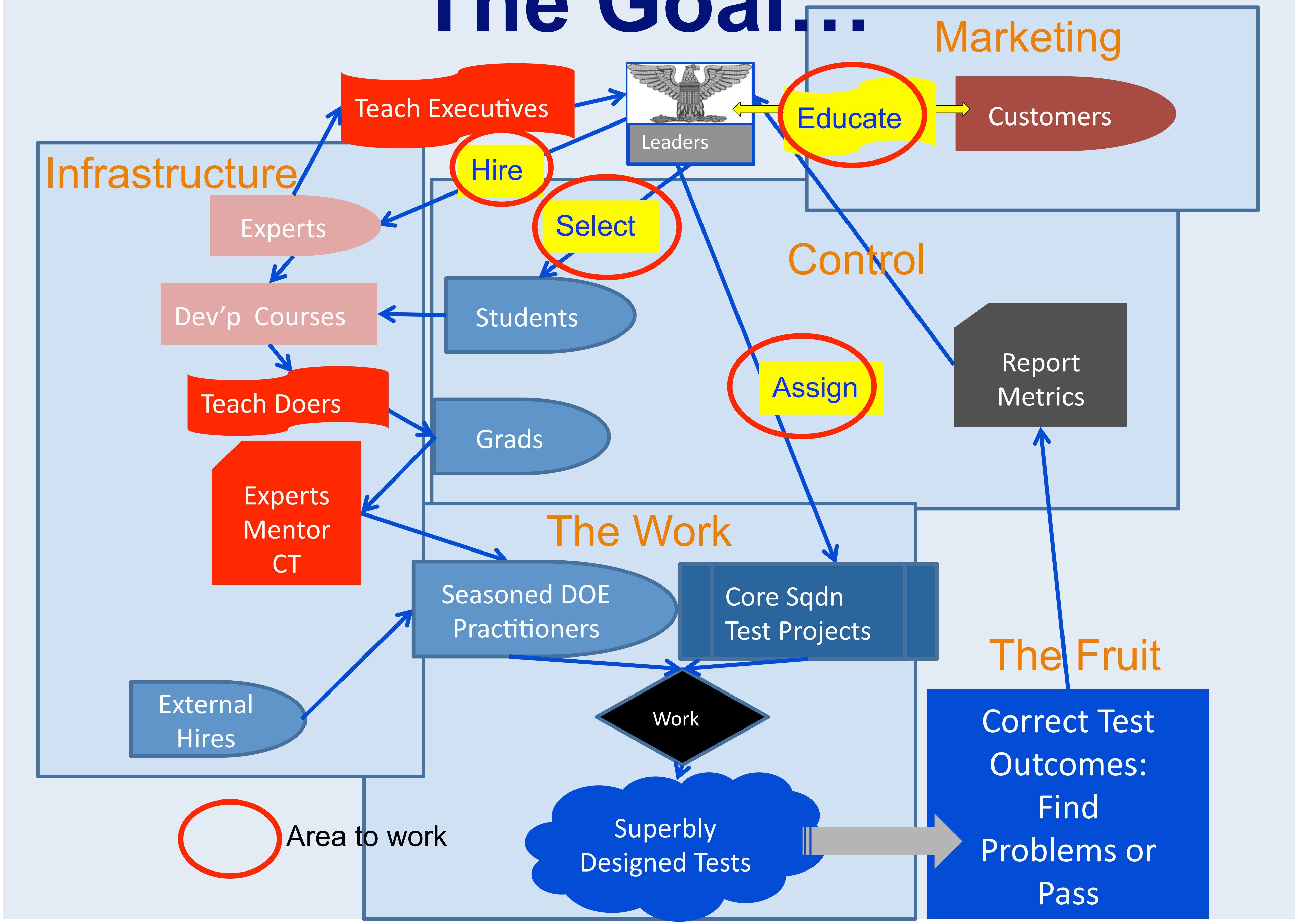


The Deeds
% Designed



The Doers

The Goal…


Marketing
Teach Executives
Leaders
Educate
Customers
Infrastructure
Hire
Experts
Select
Control
Dev'p Courses
Students
Report
Metrics
Assign
Teach Doers
Grads
Experts
Mentor
CT
The Work
Seasoned DOE
Practitioners
Core Sqdn
Test Projects
The Fruit
External
Hires
Work
Correct Test
Outcomes:
Find
Problems or
Pass
Superbly
Designed Tests
Area to work

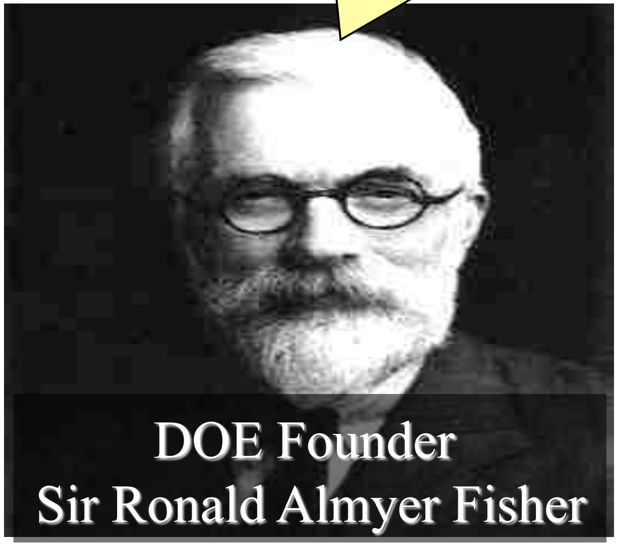
In Memorium R.A. Fisher

- Principles of DOE
  - <Orthogonality>
  - Randomization
  - Replication
  - Local Control of Error

“No aphorism is more frequently repeated in connection with field trials, than that we must ask Nature few questions, or, ideally, one question at a time. The writer is convinced that this view is wholly mistaken. Nature, he suggests, will best respond to a logical and carefully thought out questionnaire; indeed, if we ask her a single question, she will often refuse to answer until some other topic has been discussed." R. A. Fisher

“To call in the statistician after the experiment is . . . asking him to perform a postmortem examination: he may be able to say what the experiment died of.”

Address to Indian Statistical Congress, 1938.



DOE Founder
Sir Ronald Almyer Fisher

So, What's the Good News?

We Have ***Great*** Answers to ***Key*** Questions.

  - It's the way we build better tests
  - N, points, order, conclusions?
  - Uniquely answers deep and broad challenges
  - Quantify the test risks DOD incurs
  - Less-experienced testers can reliably succeed

- Small town Ga quarterback…
- A final challenge … Lead us!

George Harrison, MGen USAF (ret)

What's *Your* Method of Test?

DOE: The *Science* of Test

Questions?